Inferring key epidemiological parameters from time series and genetic data

Lucy M. Li

Feb 14, 2020

Background

Quantitative information for public health decisions during an epidemic

Questions that are asked during an epidemic may include:

1. How quickly is the infection spreading / how many people will be infected in a week?
2. How long do people remain infectious?
3. How many people would have to be vaccinated to stop the epidemic?
4. What is mortality rate?

These questions can be answerd using quantitative information that are either gathered or inferred from data. These might include:

- Basic reproductive number, R_0: The number of infections caused per infected individual in a completely susceptible population.
- Reproductive number, R: The number of infections caused per infected individual in a population that may include individuals who are not susceptible.
- Generation time, τ: The time interval between an individual becoming infected and their infector being infected.
- Duration of infection, D
- Ascertainment rate, ρ: The probability that an infected individual being diagnosed as a case. This is a function of the probability of showing symptoms given an infection, and the probability of correct diagnosis given the symptoms.
- Case-fatality rate, κ: The proportion of diagnosed individuals who pass away.

A combination of these parameter values can inform the questions above, e.g.:

1. R_0, R, τ, ρ
2. D
3. R_0, $R - R <= 1$ needed to stop an epidemic
4. ρ, κ

How to estimate epidemiological parameters?

Some parameters can be directly estimated from data. Contact tracing data provides a direct measurement of the number of infections caused by each infected individual, with some observation error e.g. missing contacts, wrong transmission link. Case-fatality rate can be calculated by dividing the number of deaths among infected patients by the total number of patients diagnosed with the infection.

Durations of infection and generation times are difficult values to obtain. The former could be estimated by using the longest generation times as a lower bound for duration of infection. Alternatively, collecting temporal samples from patients (e.g. stool, saliva, respiratory) provides information about shedding which correlates with but does not equate infectiousness. For the latter, the dates of infection cannot be directly observed and have to be inferred based on contact tracing interviews. An alternative is to use the serial interval, i.e. the time interval between symptom onset of the infector and infectee, but this can be problematic if the latent period (infected but not infectious) is long.

Ascertainment rate is a very difficult value to accurately estimate from data because the denominator (total number of infected individuals) is usually unknown Unless a seroprevalence survey is carried out across a large cross-section of the population, e.g. what they did during polio epidemics in the US in the 1950s. However, this is a value that genomic data can help to answer, because viral genomes continue to evolve as they are transmitted even in asymptomatic patients, so the genetic diversity is correlated with the total number of infections (this follows on from our previous journal club on the use of the coalescent to link population size to branching times in phylogenies).

For parameters that cannot be directly or indirectly measured, fitting mathematical models to data can help infer plausible values that explain the observed data. The two papers that we are going to discuss in today's journal club use these models to estimate R_0, ascertainment rate ρ, and other epidemiological parameters.

Besides difficulties of estimation, these paramters are population averages and do not capture variance. Large variance around reproductive numbers R can lead to superspreading events such as during the SARS outbreak when one individual caused 100+ secondary infections. Superspreading can lead to large outbreaks even if R is small if multiple superspreading events happen. Thus, taking into account the variance is also needed to accurately estimate these epidemiological parameters.

Outline

The JC will be structured as follows:

1. Introduction to compartmental and renewal models.
2. Estimation of epidemiological parameters of interest using these models.
3. Sources of uncertainty in parameter estimates.

And if there is time:

4. Fitting these models to phylogenies inferred from genomic data.

Journal club part I

Introduction to compartmental and renewal models.

Compartmental models describe groups of the population in terms of their natural history and transitions between groups. They are usually written as ordinary differential equations, as was the case in [Read2020.01.23.20018549].

The model they used was a metapopulation SEIR model. A normal SEIR model looks like this:

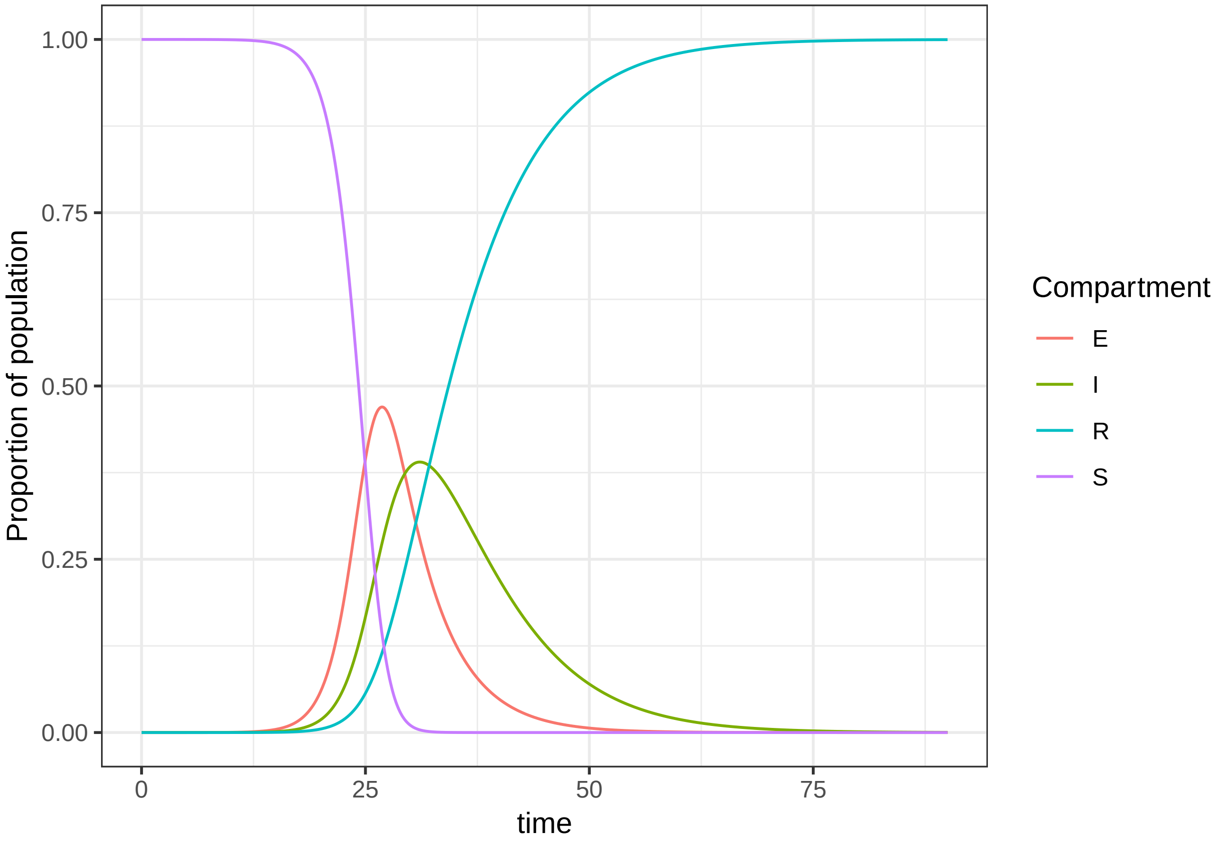
$$\begin{aligned}\frac{dS}{dt} &= -\beta IS \\ \frac{dE}{dt} &= \beta IS - \alpha E \\ \frac{dI}{dt} &= \alpha E - \gamma I \\ \frac{dR}{dt} &= \gamma I\end{aligned}$$

This assumes:

- Random mixing between all members of the population

- No births or deaths during the period that is modeled
- Individuals remain latently infected for a period of time before becoming infectious
- Individuals cannot be reinfected after clearing the infection
- Constant rates of transmission, becoming infectious, and recovering (or death)

```r
## Load deSolve package
library(deSolve)

## Create an SEIR function
seir <- function(time, state, parameters) {

  with(as.list(c(state, parameters)), {

    dS <- -beta * S * I
    dE <- beta * S * I - alpha * E
    dI <-  alpha * E - gamma * I
    dR <- gamma * I

    return(list(c(dS, dE, dI, dR)))
  })
}

### Set parameters
## Proportion in each compartment: Susceptible 0.999999, Exposed 0, Infected 0.000001, Recovered 0
init <- c(S = 1-1e-6, E=0, I = 1e-6, R = 0.0)
## alpha: becoming infectious parameter; beta: transmission parameter; gamma: recovery/removal paramete
parameters <- c(alpha=0.2, beta = 2.4247, gamma = 0.14286)
## Time frame
times <- seq(0, 90, by = .1)

## Solve using ode (General Solver for Ordinary Differential Equations)
out <- ode(y = init, times = times, func = seir, parms = parameters)
## change to data frame
out <- as.data.frame(out)
ggplot(out %>% pivot_longer(-time, names_to="Compartment")) +
  theme_bw() +
  geom_line(aes(x=time, y=value, color=Compartment)) +
  ylab("Proportion of population")
```



By estimating the parameters of this model, R_0 can be calculated using the next generation matrix method. The next generation matrix (NGM) is defined as FV^{-1} where F is the matrix that records the rates of entering one of the infected compartments (E or I in this case), and V^{-1} is the rate of leaving of the infected compartments. R_0 is the largest eigenvalue of the NGM:

$$R_0 = \rho(FV^{-1}).$$

$$F = \begin{pmatrix} \beta & 0 \\ 0 & 0 \end{pmatrix}$$

$$V = \begin{pmatrix} 0 & \alpha \\ \gamma & -\alpha \end{pmatrix}$$

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

Unfortunately the authors of [Read2020.01.23.20018549] did not provide details of the metapopulation model they used. One potential formulation is:

$$\begin{aligned}
\frac{dS_i}{dt} &= -\beta I_i S_i + \frac{S_i}{N_i}\sum_j M_{ij} \\
\frac{dE_i}{dt} &= \beta I_i S_i - \alpha E_i + \frac{E_i}{N_i}\sum_j M_{ij} \\
\frac{dI_i}{dt} &= \alpha E_i - \gamma I_i + \frac{I_i}{N_i}\sum_j M_{ij} \\
\frac{dR_i}{dt} &= \gamma I_i + \frac{R_i}{N_i}\sum_j M_{ij}
\end{aligned}$$

where M is a symmetric matrix that describes the net migration rate between major cities in China. This assumes all individuals are equally likely to migrate.

Alternatively, the transmission rate β can be defined between and within cities:

$$\begin{aligned}
\frac{dS_i}{dt} &= -\sum_j (\beta_{ij} I_j) S_i \\
\frac{dE_i}{dt} &= \sum_j (\beta_{ij} I_j) I_i S_i - \alpha E_i \\
\frac{dI_i}{dt} &= \alpha E_i - \gamma I_i \\
\frac{dR_i}{dt} &= \gamma I_i
\end{aligned}$$

where β_{ij} is the transmission rate from infectious individuals in city j to susceptible individuals i.

Stochastic versions of compartmental models can better capture random effects, which are especially important at the start and end of epidemics when the number of infected individuals is small. There are many ways to formulate stochastic models. For compartmental models, they are usually forumlated as State Space Models (SSMs), a type of Hidden Markov Model (HMM).

In the case of the second paper (Li et al.), the stochastic renewal model they used only directly modeled the number of individuals in the 'Infectious' class.

$$\begin{aligned}
I(t) &= \int_0^\infty I(t-\tau)\beta(\tau)d\tau \\
R &= \int \beta(\tau)d\tau
\end{aligned}$$

We can define a generation time probability distribution $\omega(\tau)$:

$$\int \omega(\tau)d\tau = \frac{1}{R}\int \beta(\tau)d\tau$$

if we assume the rate of transmission per unit of time is constant. We can re-write the equation for $I(t)$ above:

$$\begin{aligned}
\beta(\tau) &= R \cdot \omega(\tau) \\
I(t) &= R\int_0^\infty I(t-\tau)\omega(\tau)d\tau \\
R &= \frac{I(t)}{\int_0^\infty I(t-\tau)\omega(\tau)d\tau}
\end{aligned}$$

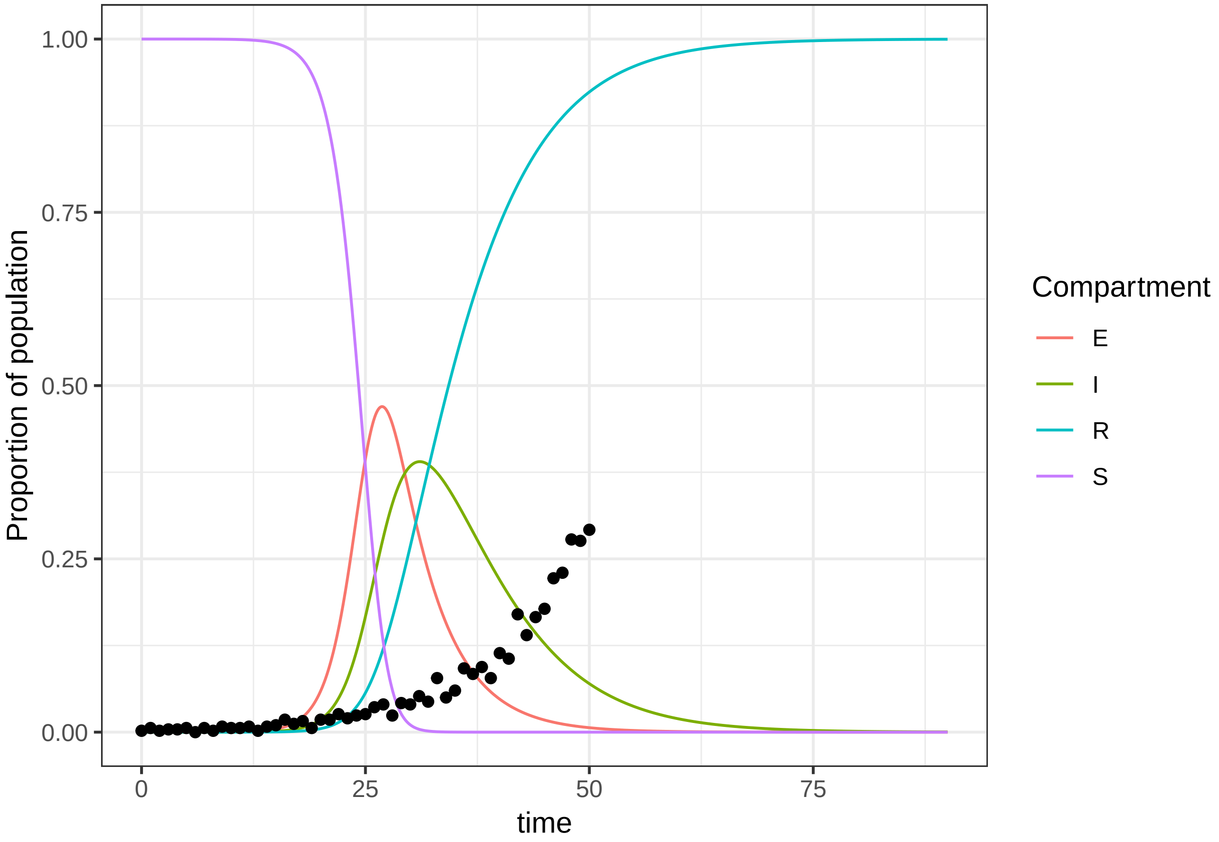
During the early stages of an epidemic, number of infected individuals grows exponentially. We can parameterize the renewal equation above using an exponential model and calculate R_0 as:

$$\begin{aligned} I(t) &= I(t_0)\exp^{r(t-t_0)} \\ \text{Let } I(t_0) &= 1 \\ R_0 &= \frac{1}{\int_0^\infty e^{-rt}\omega(\tau)d\tau} \end{aligned}$$

Estimation of epidemiological parameters of interest using these models.

In both cases, the models were fit to incidence time series (i.e. the epi curve). This means the set of parameters values that best describes the observed time series.

```r
simdata <- data.frame(time=seq(0, 50)) %>% mutate(data=exp(.1 * time) %>% sapply(function (x) rpois(1, 
ggplot(out %>% pivot_longer(-time, names_to="Compartment")) +
  theme_bw() +
  geom_line(aes(x=time, y=value, color=Compartment)) +
  geom_point(data=simdata, aes(x=time, y=data)) +
  ylab("Proportion of population")
```



Sources of uncertainty in parameter estimates and heterogeneity in transmission.

The quantities above are averaged across all individuals in a population. The values can be highly variable, so it is also useful to describe the distribution of these values:

- $\nu \sim \phi(\nu)$, where ν is the reproductive number of an individual, and $\phi(\nu)$ is the probability distribution of the reproductive number, also known as the offspring distribution in population genetic terms.

Also, what do the dates in the data mean? Dates of onset? Dates of collection?

Journal club part II

Fitting epidemiological models to phylogenies inferred from genomic data.

References

Li, Qun, Xuhua Guan, Peng Wu, Xiaoye Wang, Lei Zhou, Yeqing Tong, Ruiqi Ren, et al.“Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia.” *New England Journal of Medicine* 0 (0): null. https://doi.org/10.1056/NEJMoa2001316.