A note on observational processes in epidemic models

Sang Woo Park and Benjamin M. Bolker

July 22, 2019

1 Introduction

Mechanistic analyses of epidemic time series allow us to infer the underlying transmission mechanism, estimate biologically relevant parameters, and predict the course of an epidemic (Bretó et al., 2009). In order to make a precise and accurate inference, disease modelers often focus on developing more realistic process models. For example, the same London measles time series from the prevaccination era has been analyzed using multiple models – these models account for time-varying transmission rates (Fine and Clarkson, 1982), realistic age structure (Schenzle, 1984), metapopulation structure (Xia et al., 2004), continuous-time infection process (Cauchemez and Ferguson, 2008), and extra-demographic variability (He et al., 2009).

Despite the amount of effort put into developing *better* process models, disease modelers often neglect details of the observation process associated with disease case reports. Most disease models assume that new cases are reported instantaneously when an individual is infected (e.g., Martinez et al. (2016); Kennedy et al. (2018); Pons-Salort and Grassly (2018)) or develops symptom; one exception is the model by (He et al., 2009), which assumes that infections are counted upon recovery (because diagnosed cases are put to bed rest and are effectively no longer infectious).

Here, we use a simple Susceptible-Infected-Removed (SIR) model to study how assumptions about the underlying observation process affect parameter estimates of the SIR model. We show that making wrong assumptions about the timing of incidence reports can lead to biases in parameter estimates and narrow confidence intervals.

2 Methods

The Susceptible-Infected-Removed (SIR) model describes how disease spreads in a homogeneous population:

$$\begin{aligned}\frac{dS}{dt} &= -\beta S \frac{I}{N} \\ \frac{dI}{dt} &= \beta S \frac{I}{N} - \gamma I \\ \frac{dR}{dt} &= \gamma I,\end{aligned} \tag{1}$$

where β is the contact rate, γ is the recovery rate, and $N = S + I + R$ is the total population size. We define incidence at time t as the number of newly infected individuals that are infected between time $t - \Delta t$ and time t, where Δt is the reporting time step. We expect infected individuals to report their infection some time after their infection; the number of reported cases defines the *observed* incidence.

For brevity, we consider two extreme cases: individuals instantaneously report their infection when they become infected or when they recover. The observed incidence measured upon infection ($i_1(t)$) can be defined by the integral:

$$i_1(t) = \int_{t-\Delta t}^{t} \beta S \frac{I}{N} dt, \tag{2}$$

where Δt is the reporting time step. Alternatively, we can keep track of cumulative incidence, C, by adding a new state variable described by $dC/dt = \beta SI/N$ and taking the difference between the two consecutive reporting periods: $i_1(t) = C(t) - C(t - \Delta t)$. Likewise, incidence measured upon recovery ($i_2(t)$) can be defined by the integral:

$$i_2(t) = \int_{t-\Delta t}^{t} \gamma I dt, \tag{3}$$

or by the consecutive difference in the cumulative number of recovered cases: $i_2(t) = R(t) - R(t - \Delta t)$. Finally, we model under-reporting using a negative binomial distribution with mean $\rho i_1(t)$ or $\rho i_2(t)$, where ρ is the reporting rate, and an over dispersion parameter θ. For convenience, we will refer to these two negative binomial models as infection model and recovery model hereafter; likewise, we will refer to epidemic time series generated from these models as infection time series and recovery time series.

In this study, we focus on estimating 5 parameters: the basic reproductive number $\mathcal{R}_0 = \beta/\gamma$, mean infectious period $1/\gamma$, reporting rate ρ, an over dispersion parameter θ and the initial proportion of the infected individuals i_0. Initial proportion of susceptible individuals is assumed to be $1 - i_0$. The total population size N is assumed to be known.

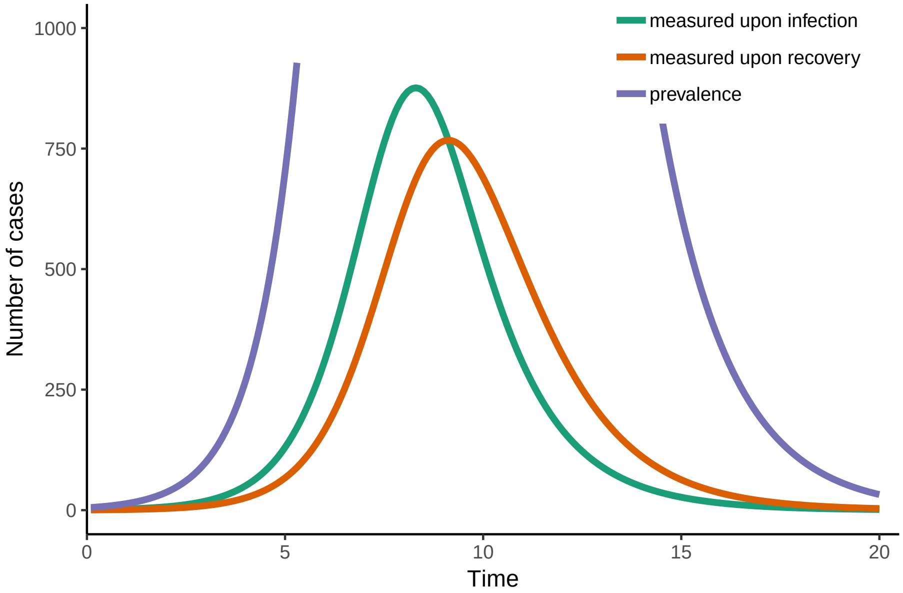

Figure 1: **A comparison of incidence measured at two different time points in infection.** We used the following parameters to simulate the SIR model and generate the figure: $\mathcal{R}_0 = 2$, $1/\gamma = 1$ time units, $N = 1 \times 10^5$, $i0 = 1 \times 10^{-4}$, and $\Delta t = 0.1$ time units.

3 Results

In order to introduce the problem, we first compare the dynamics of observed incidence, $i_1(t)$ and $i_2(t)$, (Figure 1). Lags in reporting time delay the observed epidemic peak timing and reduce the size of the peak. However, these differences in the reporting time have small effect to the overall shape of the epidemic curve. In the presence of observation and process error, we would not expect to be able to distinguish between the two reporting processes.

Note that incidence measured at these two periods are different from prevalence, $I(t)$, which is defined as the number of *currently* infected individuals; the dynamics of observed incidence depend on the reporting time step (because the sum of true incidence is equal to the final size of an epidemic), whereas those of prevalence do not. We expect the dynamics of incidence and prevalence to be similar only when the reporting time step is equal to the disease generation time. While it is uncommon, some models do not make a clear distinction between the two.

In order to understand how assumptions about the timing of case reporting affect parameter estimates of the SIR model, we simulate observed epidemic curves (i.e., infection time series and recovery time series), measured upon both infection and recovery, 100 times and fit both models (i.e., infection model

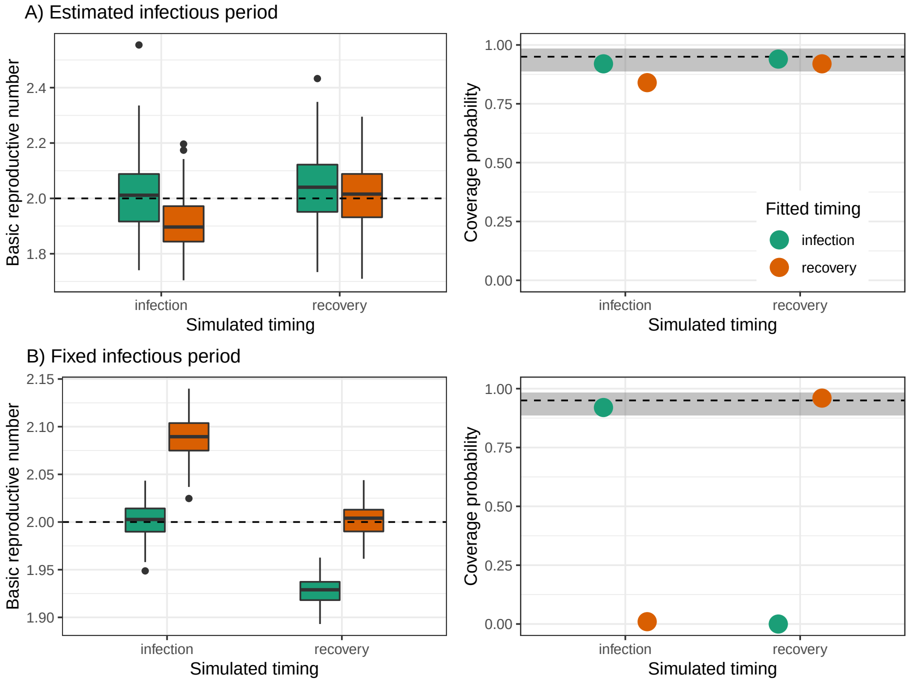

Figure 2: **A comparison of incidence measured at two different time points in infection.** We used the following parameters to simulate the SIR model: $\mathcal{R}_0 = 2$, $1/\gamma = 1$ time units, $N = 1 \times 10^5$, $i0 = 1 \times 10^{-4}$, $\theta = 10$, and $\Delta t = 0.1$ time units.

and recovery model) to each time series. We compare the estimates of the basic reproductive number $\mathcal{R}_0$, and its coverage probability, defined as the proportion of confidence intervals that contain the true value (95% confidence interval is expected to contain the true value 95% of the time by definition). We summarize the results in Figure 2.

When we try to estimate all 5 parameters, fitting the recovery model to infection time series underestimates the basic reproduction number and gives a slightly low coverage (Figure 2A). Fitting the infection model to recovery time series slightly overestimates the basic reproduction number but gives good coverage. Fitting the correct models to their corresponding time series gives unbiased estimates and good coverage.

Disease modelers often assume that the mean infectious period of a disease is known and focus on estimating the basic reproduction number (or the transmission rate). When we assume that the mean infectious period is known and try to estimate the remaining 4 parameters of the SIR model, fitting the wrong model results in clearer bias and lower coverage (Figure 2B); however, bias from fitting the wrong model changes direction (e.g., fitting the infection model to recovery model overestiamtes the reproduction number instead).

Differences in the direction of the bias can be explaind by the estimates of the exponential growth rate ($r = \beta - \gamma$) and the mean infectious period ($1/\gamma$). When we fit the recovery model to infection time series, the model overestimates the initial growth rate in order to match the higher epidemic peak of the infection time series. When the mean infectious period is fixed, higher growth rate translates to higher basic reproduction number as we see in Figure 2B. When we allow the mean infectious period to vary, the model underestimates the mean infectious period (high γ), which also decreases the estimate of the basic reproduction number. Likewise, fitting the infection model to recovery time series overestimates the mean infectious period (low γ), increasing the estimate of the basic reproduction number; when the mean infectious period is fixed, low growth rate estimates (to match the lower epidemic peak of the recovery time series) decrease the estimate of the basic reproduction number.

4 Discussion

Mathematical modeling of infectious disease outbreaks helps us understand how disease spreads in a population. Disease modelers often put a lot of effort into capturing the transmission mechanism but make naive assumptions about how cases are reported. Here, we used a simple deterministic SIR model to show that making wrong assumptions about observational processes in epidemic models can give biased estimates of the basic reproduction number and narrow confidence intervals.

We assumed that newly infected cases are reported instantaneously when individuals become infected or when they recover. Although neither of these assumptions are realistic, they are broadly consistent with the assumptions of other state-of-art epidemic models. More realistic models may distinguish reported and unreported cases by adding a new state variable or by modeling an explicit delay kernel in reporting time.

While we have only considered observation processes associated with incidence reporting in this study, it is equally as important to consider observation processes in mortality reporting. In particular, many disease modelers have tried to infer underlying transmission mechanisms from historical mortality data but assumed that deaths are recorded as soon as individuals die. This assumption does not account for the possibility that delays in reporting time of deaths can change the shape of an epidemic curve: longer delays are expected to decrease the size of an epidemic peak and delay the timing of the peak (Figure 1). These assumptions may have important effects on their conclusions about the underlying transmission mechanism.

Our study shows that seemingly irrelevant changes in the assumptions of an epidemic model affects the inference of infectious disease transmission. We caution disease modelers to be more mindful about their decisions in developing epidemic models and the implications of their model assumptions.

References

Bretó, C., D. He, E. L. Ionides, A. A. King, et al. (2009). Time series analysis via mechanistic models. *The Annals of Applied Statistics 3*(1), 319–348.

Cauchemez, S. and N. M. Ferguson (2008). Likelihood-based estimation of continuous-time epidemic models from time-series data: application to measles transmission in London. *Journal of the Royal Society Interface 5*(25), 885–897.

Fine, P. E. and J. A. Clarkson (1982). Measles in England and Wales—I: an analysis of factors underlying seasonal patterns. *International journal of epidemiology 11*(1), 5–14.

He, D., E. L. Ionides, and A. A. King (2009). Plug-and-play inference for disease dynamics: measles in large and small populations as a case study. *Journal of the Royal Society Interface 7*(43), 271–283.

Kennedy, D. A., P. A. Dunn, and A. F. Read (2018). Modeling Marek's disease virus transmission: A framework for evaluating the impact of farming practices and evolution. *Epidemics 23*, 85–95.

Martinez, P. P., A. A. King, M. Yunus, A. Faruque, and M. Pascual (2016). Differential and enhanced response to climate forcing in diarrheal disease due to rotavirus across a megacity of the developing world. *Proceedings of the National Academy of Sciences 113*(15), 4092–4097.

Pons-Salort, M. and N. C. Grassly (2018). Serotype-specific immunity explains the incidence of diseases caused by human enteroviruses. *Science 361*(6404), 800–803.

Schenzle, D. (1984). An age-structured model of pre-and post-vaccination measles transmission. *Mathematical Medicine and Biology: A Journal of the IMA 1*(2), 169–191.

Xia, Y., O. N. Bjørnstad, and B. T. Grenfell (2004). Measles metapopulation dynamics: a gravity model for epidemiological coupling and dynamics. *The American Naturalist 164*(2), 267–281.