A note on observation processes in epidemic models

Sang Woo Park[1,2,*] Benjamin M. Bolker[2,3,4]

1 Department of Ecology and Evolutionary Biology, Princeton University, Princeton, New Jersey, USA
2 Department of Mathematics and Statistics, McMaster University, Hamilton, Ontario, Canada
3 Department of Biology, McMaster University, Hamilton, Ontario, Canada
4 Michael G. DeGroote Institute for Infectious Disease Research, McMaster University, Hamilton, Ontario, Canada

*swp2@princeton.edu

Abstract

Many disease models focus on characterizing the underlying transmission mechanism but make simple, possibly naive assumptions about how infections are reported. In this note, we use a simple deterministic Susceptible-Infected-Removed (SIR) model to compare two common assumptions about disease incidence reports: individuals can report their infection as soon as they become infected or as soon as they recover. We show that incorrect assumptions about the underlying observation processes can bias estimates of the basic reproduction number and lead to overly narrow confidence intervals.

1 Introduction

Mechanistic analyses of epidemic time series allow us to make inference about the underlying transmission mechanisms, estimate biologically relevant parameters, and predict the course of an outbreak (Bretó et al., 2009). In order to make precise and accurate inferences, disease modelers have sought to build more realistic process models. For example, the time series of reported measles cases from London in the prevaccination era has been analyzed many times, using variety of models accounting for time-varying transmission rates (Fine and Clarkson, 1982), realistic age structure (Schenzle, 1984), metapopulation structure (Xia et al., 2004), continuous-time infection processes (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 processes associated with new disease case reports (often referred to as incidence time series). Many disease models effectively 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 when an individual becomes symptomatic (e.g., Bhadra et al. (2011); King et al. (2015)); some models (e.g., Bretó et al.

(2009); He et al. (2009); Lin et al. (2016)) assume that infections are counted upon recovery (because diagnosed cases are controlled and are effectively no longer infectious).

We emphasize that incidence (i.e., the number of *newly* infected individuals) is different from prevalence (i.e., the number of *currently* infected individuals) (Bjørnstad, 2018). The dynamics of 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 similar to the disease generation time (Fine and Clarkson, 1982). While they are uncommon, some models do not make a clear distinction between prevalence and incidence (Capistrán et al., 2009; Hooker et al., 2010; Yang et al., 2013; González-Parra et al., 2014).

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

2 Methods

The Susceptible-Infected-Removed (SIR) model describes how a 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 per unit time, γ is the recovery rate per unit time, and $N = S + I + R$ is the total population size. We define *true* 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 cases to be reported some time after infection; the number of reported cases during a time period 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:

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

Equivalently, 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, the observed 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 observation error using a negative binomial distribution with a mean of either $\rho i_1(t)$ 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; similarly, we will refer to epidemic time series generated from these negative binomial models with two different means ($\rho i_1(t)$ and $\rho i_2(t)$) 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, the overdispersion parameter θ, and the initial proportion of the infected individuals i_0. The initial proportion of susceptible individuals is assumed to be $1 - i_0$. The total population size N is assumed to be known.

3 Results

Figure 1A compares the deterministic dynamics of two incidence curves, $i_1(t)$ and $i_2(t)$, and a prevalence curve $I(t)$ for $\mathcal{R}_0 = 2$. A lag in reporting time delays the timing of the observed epidemic peak and reduces the size of that peak. As $\mathcal{R}_0$ increases, the difference in the size of the peaks increases (Figure 1B) but the difference in the timing of the peaks decreases (Figure 1C). For example, when $\mathcal{R}_0 = 5$, such delay in the reporting of new cases can reduce the size of the observed epidemic peak by almost 50%. Note that $i_2(t)$ essentially assumes that the amount of time between when an individual is infected and when an individual reports the infection is exponentially distributed; a fixed delay in reporting time will not change the shape of an epidemic curve.

For small values of $\mathcal{R}_0$, these differences in the reporting time have little effect on the overall shape of the epidemic curve. In the presence of observation and process error, we do not expect to be able to distinguish between the two reporting processes based on the time series alone (and we rarely know *a priori* the delay distribution between when an individual is infected and when that case is reported). Therefore, one might naively expect assumptions about the timing of case reporting to have negligible effect on inference.

In order to understand how assumptions about the timing of case reporting affect parameter estimates of the SIR model, we simulate epidemic time series (infection time series and recovery time series) 100 times with $\mathcal{R}_0 = 2$ and fit both infection and recovery models to each time series. We compare the estimates of the basic reproductive number $\mathcal{R}_0$, and the coverage of our confidence intervals, 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). Figure 2 summarizes the results.

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. We expect fitting incorrect models to give more biased estimates when $\mathcal{R}_0$ is higher because the differences in the size of observed epidemic peaks become greater. Fitting the correct model gives unbiased estimates and

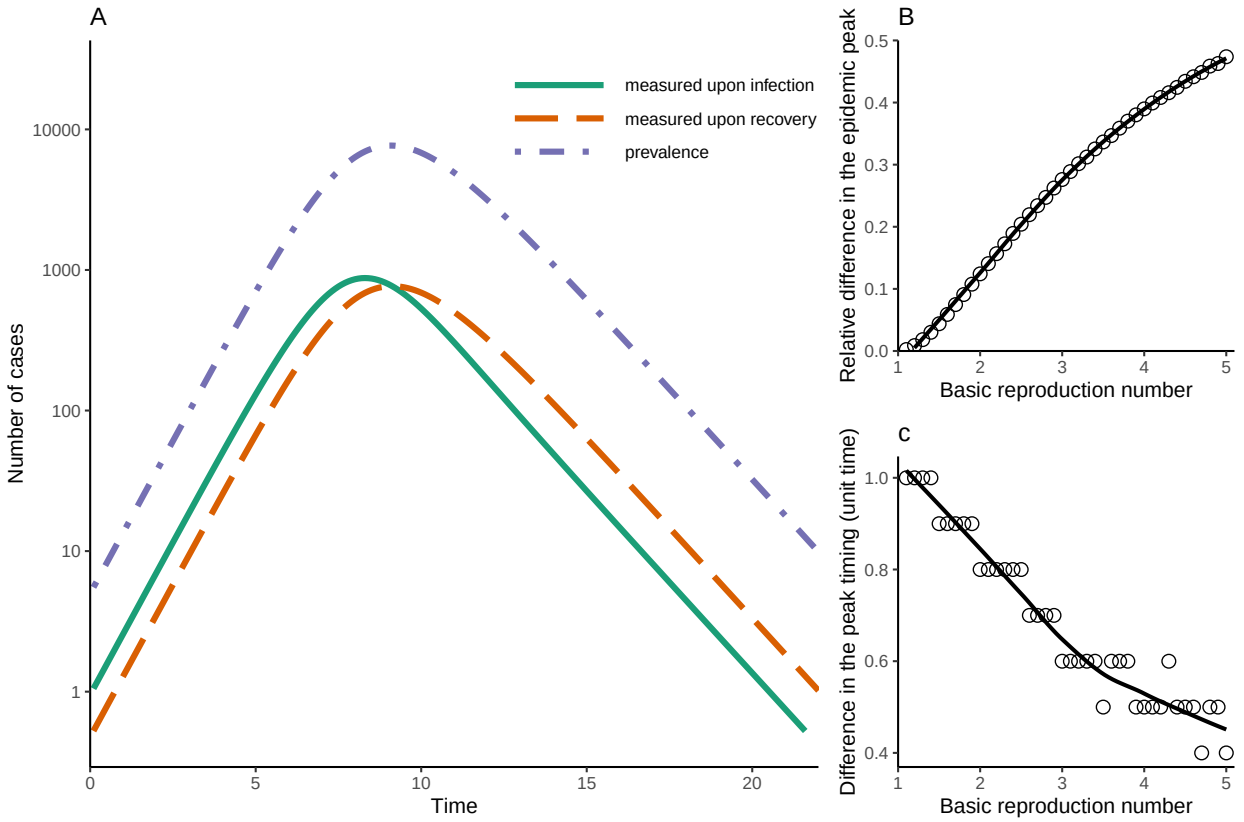

Figure 1: **A comparison of incidence measured at two different time points in infection.** (A) A deterministic simulation of the SIR model using the following parameters: $\mathcal{R}_0 = 2$, $1/\gamma = 1$ time units, $N = 1 \times 10^5$, $i_0 = 1 \times 10^{-4}$, and $\Delta t = 0.1$ time units. (B-C) Effects of $\mathcal{R}_0$ on the relative difference in the size of the epidemic peak $(1 - [\max i_2(t)]/[\max i_1(t)])$ and the difference in the peak timing $(\hat{t}_2 - \hat{t}_1$ where $i_k(\hat{t}_k) = \max i_k(t)$ for $k = 1, 2)$ due to delays in reporting time. Open circles: simulation results of the deterministic SIR model. Solid lines: locally estimated scatterplot smoothing (LOESS) curves fitted to simulation results. Remaining parameters are held constant ($1/\gamma = 1$ time units, $N = 1 \times 10^5$, $i_0 = 1 \times 10^{-4}$, and $\Delta t = 0.1$ time units) throughout simulations.

good coverage.

Disease modelers often assume that the mean infectious period of a disease is known and focus on estimating the basic reproduction number (e.g., Hooker et al. (2010); Lin et al. (2016); Pons-Salort and Grassly (2018)). When we assume that the true value of the mean infectious period is known and try to estimate the remaining 4 parameters of the SIR model (Figure 2B), fitting incorrect models results in a clearer bias (in opposite directions) and a much lower coverage (cf. Elderd et al. (2006)).

Differences in the direction of the bias can be explained by the estimates of the exponential growth rate ($r = \beta - \gamma$) and the mean infectious period ($1/\gamma$). In general, we expect delays in

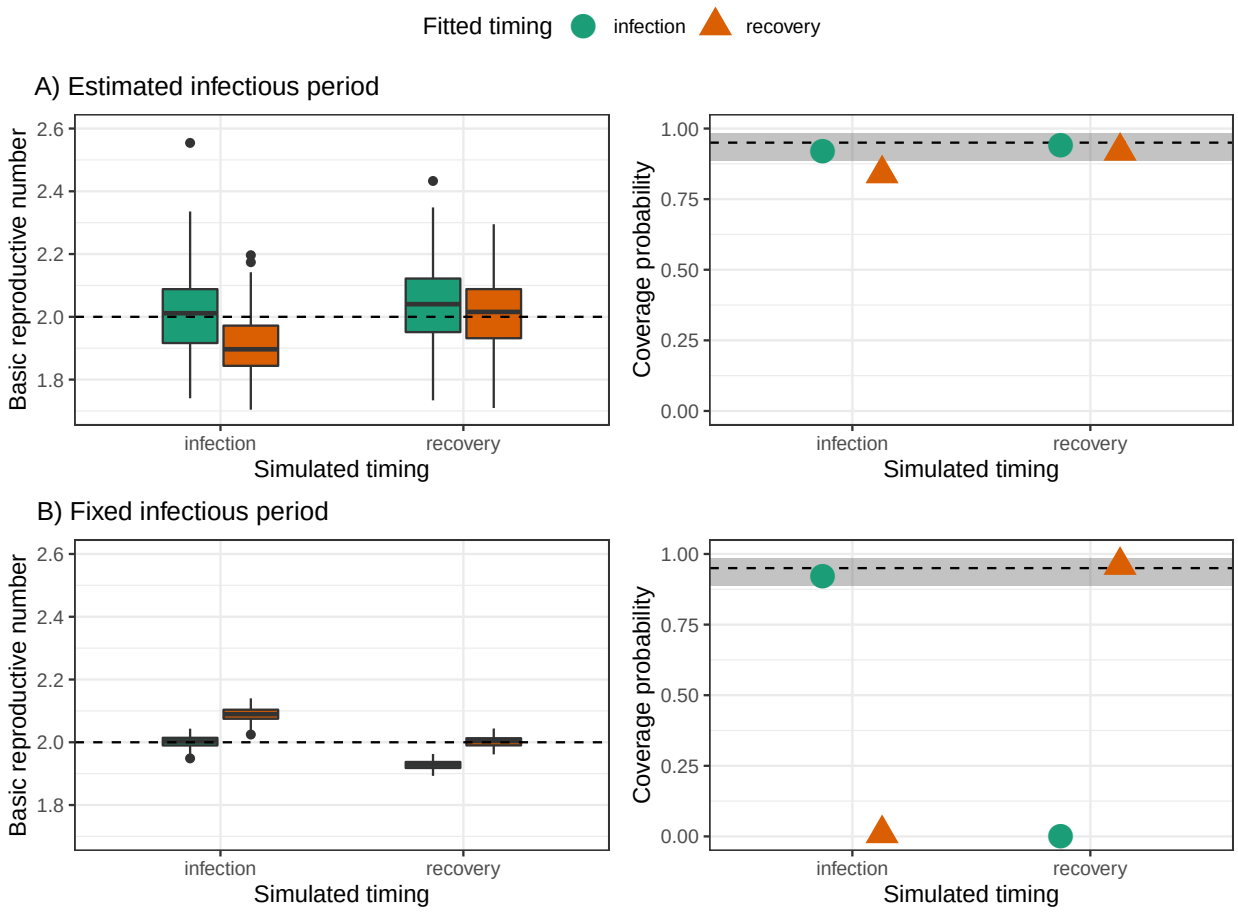

Figure 2: **A comparison of incidence measured at two different time points in infection.** We simulate infection time series and recovery time series 100 times using the following parameters: $\mathcal{R}_0 = 2$, $1/\gamma = 1$ time units, $N = 1 \times 10^5$, $i_0 = 1 \times 10^{-4}$, $\rho = 0.5$, $\theta = 10$, and $\Delta t = 0.1$ time units. For each simulation, we fit the infection model and the recovery model by (A) estimating the mean infectious period and (B) assuming that the mean infectious period is known. Coverage probabilities represent the proportion of confidence intervals that contain the true value of the basic reproductive number ($\mathcal{R}_0$).

observation processes to make the observed epidemic time series last longer and have smaller peaks (Figure 1). When we fit the recovery model to an infection time series, the model overestimates the initial growth rate in order to match the bigger (and faster) epidemic peak of the infection time series. When the mean infectious period is fixed, higher growth rate translates to higher basic reproduction number, as Figure 2B shows. When we allow the mean infectious period to vary, the model still overestimates the growth rate but also underestimates the mean infectious period (high γ), which decreases the overall estimate of the basic reproduction number. Similarly, fitting the infection model to a recovery time series underestimates the growth rate to match the smaller (and slower) epidemic peak; this underestimates the basic reproduction number when the mean infectious period is fixed. When we allow the mean infectious period to vary, we overestimate the mean infectious period, which in turn increases the estimate of the basic reproduction number.

4 Discussion

Mathematical modeling of infectious disease outbreaks helps us understand how disease spreads in a population; however, epidemic models often make simple assumptions about how cases are reported. We used a deterministic SIR model to show that delays in case reports affect the observed shape of an epidemic curve and the estimates of the basic reproduction number. Even when the basic reproduction number is small (e.g., $\mathcal{R}_0 = 2$), fitting incorrect models can introduce small bias and give narrow confidence intervals.

We compared two scenarios in which newly infected cases are reported instantaneously (1) when individuals become infected or (2) when they recover. Although neither of these assumptions is realistic, many epidemic models still rely on these assumptions (see Introduction). More realistic models may distinguish reported and unreported (i.e., identified and unidentified) cases by adding new state variables (Browne et al., 2015; Webb et al., 2015) or by modeling an explicit delay distribution in reporting time (Harris, 1990; Ferguson et al., 2001; Goldstein et al., 2009; Ster et al., 2009; Birrell et al., 2011; Funk et al., 2018).

We considered observation processes associated with incidence reporting; however, we expect observation processes to be just as (if not more) important in analyzing mortality data. Many disease modelers have tried to infer underlying transmission mechanisms from historical mortality data but assumed that individual deaths are recorded as soon as individuals die (He et al., 2013; Didelot et al., 2017; Dean et al., 2018); this includes the classic work by Kermack and McKendrick (1927) who approximated the reported *number* of deaths per week from plague with instantaneous death *rates* (dR/dt). These frameworks do not account for the possibility that a delay in reporting of deaths can change the shape of an epidemic curve – delays in case reports can decrease the size of the observed epidemic peak and delay the observed timing of the peak (Figure 1).

Distributions of reporting time delays implicitly depend on the latent and infectious period distributions. For example, when cases are reported at the end of infectious periods (Bretó et al., 2009; He et al., 2009; Lin et al., 2016), the delay distribution is equivalent to the convolution of the latent and infectious period distributions. These distributions also

play important roles in shaping the epidemiological dynamics, including the probability and size of an outbreak (Anderson and Watson, 1980), the stability and persistence of a system (Keeling and Grenfell, 1998; Lloyd, 2001a,b), and the occurrence of dynamical transitions (Krylova and Earn, 2013). Assumptions about latent and infectious period distributions can also have large effects on the estimates of the basic reproductive number (Wearing et al., 2005).

Here, we assumed that the underlying transmission process is deterministic; this assumes that all error can be explained by observation errors alone. We chose to study a deterministic model SIR for computational efficiency; we do not recommend using deterministic models for real outbreak analyses. Ignoring process errors (i.e., stochasticity in the transmission process) can lead to overly confident results (King et al., 2015). Using stochastic models may give better coverage probabilities even when wrong observation models are used. Nonetheless, misspecifying the observation model may still affect conclusions from stochastic models and introduce bias.

All statistical models rely on simplifying assumptions. Given a range of possible assumptions, modelers have to decide which aspects of a system to capture at the cost of simplifying other aspects. However, some assumptions are often taken for granted. Epidemic modelers tend to focus on assumptions about the transmission process; processes through which cases are reported are often overlooked and simplified. Our study shows that a seemingly negligible change in the assumptions of an epidemic model can affect the inference of infectious disease transmission. Although the amount of bias introduced from the misspecification of the observation model can be small (e.g., less than 5% in our examples), modelers trying to make serious predictions may want to control for potential biases. We caution disease modelers to be more careful about their modeling decisions as they may have unexpected consequences on the inference.

Acknowledgements

We thank David Earn for providing helpful comments on the manuscript. BMB is supported by an NSERC Discovery grant.

References

Anderson, D. and R. Watson (1980). On the spread of a disease with gamma distributed latent and infectious periods. *Biometrika 67*(1), 191–198.

Bhadra, A., E. L. Ionides, K. Laneri, M. Pascual, M. Bouma, and R. C. Dhiman (2011). Malaria in Northwest India: Data analysis via partially observed stochastic differential equation models driven by Lévy noise. *Journal of the American Statistical Association 106*(494), 440–451.

Birrell, P. J., G. Ketsetzis, N. J. Gay, B. S. Cooper, A. M. Presanis, R. J. Harris, A. Charlett, X.-S. Zhang, P. J. White, R. G. Pebody, et al. (2011). Bayesian modeling to unmask and

predict influenza A/H1N1pdm dynamics in London. *Proceedings of the National Academy of Sciences 108*(45), 18238–18243.

Bjørnstad, O. N. (2018). *Epidemics: models and data using R*. Springer.

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.

Browne, C., H. Gulbudak, and G. Webb (2015). Modeling contact tracing in outbreaks with application to Ebola. *Journal of Theoretical Biology 384*, 33–49.

Capistrán, M. A., M. A. Moreles, and B. Lara (2009). Parameter estimation of some epidemic models. The case of recurrent epidemics caused by respiratory syncytial virus. *Bulletin of Mathematical Biology 71*(8), 1890.

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.

Dean, K. R., F. Krauer, L. Walløe, O. C. Lingjærde, B. Bramanti, N. C. Stenseth, and B. V. Schmid (2018). Human ectoparasites and the spread of plague in Europe during the Second Pandemic. *Proceedings of the National Academy of Sciences 115*(6), 1304–1309.

Didelot, X., L. K. Whittles, and I. Hall (2017). Model-based analysis of an outbreak of bubonic plague in Cairo in 1801. *Journal of The Royal Society Interface 14*(131), 20170160.

Elderd, B. D., V. M. Dukic, and G. Dwyer (2006). Uncertainty in predictions of disease spread and public health responses to bioterrorism and emerging diseases. *Proceedings of the National Academy of Sciences 103*(42), 15693–15697.

Ferguson, N. M., C. A. Donnelly, and R. M. Anderson (2001). The foot-and-mouth epidemic in Great Britain: pattern of spread and impact of interventions. *Science 292*(5519), 1155–1160.

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.

Funk, S., A. Camacho, A. J. Kucharski, R. M. Eggo, and W. J. Edmunds (2018). Real-time forecasting of infectious disease dynamics with a stochastic semi-mechanistic model. *Epidemics 22*, 56–61.

Goldstein, E., J. Dushoff, J. Ma, J. B. Plotkin, D. J. Earn, and M. Lipsitch (2009). Reconstructing influenza incidence by deconvolution of daily mortality time series. *Proceedings of the National Academy of Sciences 106*(51), 21825–21829.

González-Parra, G., A. J. Arenas, and B. M. Chen-Charpentier (2014). A fractional order epidemic model for the simulation of outbreaks of influenza A (H1N1). *Mathematical Methods in the Applied Sciences 37*(15), 2218–2226.

Harris, J. E. (1990). Reporting delays and the incidence of AIDS. *Journal of the American Statistical Association 85*(412), 915–924.

He, D., J. Dushoff, T. Day, J. Ma, and D. J. Earn (2013). Inferring the causes of the three waves of the 1918 influenza pandemic in England and Wales. *Proceedings of the Royal Society B: Biological Sciences 280*(1766), 20131345.

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.

Hooker, G., S. P. Ellner, L. D. V. Roditi, and D. J. Earn (2010). Parameterizing state–space models for infectious disease dynamics by generalized profiling: measles in Ontario. *Journal of The Royal Society Interface 8*(60), 961–974.

Keeling, M. and B. T. Grenfell (1998). Effect of variability in infection period on the persistence and spatial spread of infectious diseases. *Mathematical Biosciences 147*(2), 207–226.

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.

Kermack, W. O. and A. G. McKendrick (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of A: Mathematical, Physical and Engineering Sciences 115*(772), 700–721.

King, A. A., M. Domenech de Cellès, F. M. Magpantay, and P. Rohani (2015). Avoidable errors in the modelling of outbreaks of emerging pathogens, with special reference to Ebola. *Proceedings of the Royal Society B: Biological Sciences 282*(1806), 20150347.

Krylova, O. and D. J. Earn (2013). Effects of the infectious period distribution on predicted transitions in childhood disease dynamics. *Journal of The Royal Society Interface 10*(84), 20130098.

Lin, Q., Z. Lin, A. P. Chiu, and D. He (2016). Seasonality of influenza A (H7N9) virus in China-—fitting simple epidemic models to human cases. *PLoS one 11*(3), e0151333.

Lloyd, A. L. (2001a). Destabilization of epidemic models with the inclusion of realistic distributions of infectious periods. *Proceedings of the Royal Society of B: Biological Sciences 268*(1470), 985–993.

Lloyd, A. L. (2001b). Realistic distributions of infectious periods in epidemic models: changing patterns of persistence and dynamics. *Theoretical Population Biology 60*(1), 59–71.

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.

Ster, I. C., B. K. Singh, and N. M. Ferguson (2009). Epidemiological inference for partially observed epidemics: the example of the 2001 foot and mouth epidemic in Great Britain. *Epidemics 1*(1), 21–34.

Wearing, H. J., P. Rohani, and M. J. Keeling (2005). Appropriate models for the management of infectious diseases. *PLoS medicine 2*(7).

Webb, G., C. Browne, X. Huo, O. Seydi, M. Seydi, and P. Magal (2015). A model of the 2014 Ebola epidemic in West Africa with contact tracing. *PLoS currents 7*.

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.

Yang, J.-Y., Y. Chen, and F.-Q. Zhang (2013). Stability analysis and optimal control of a hand-foot-mouth disease (HFMD) model. *Journal of Applied Mathematics and Computing 41*(1-2), 99–117.