A note on observation processes in epidemic models

Sang Woo Park and Benjamin M. Bolker

July 31, 2019

Abstract

Epidemic models provide valuable tools for understanding how a disease spreads in a population. Many epidemic models focus on characterizing the underlying transmission mechanism but make 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 making wrong assumptions about the underlying observation processes can result in biased estimates of the basic reproduction number and narrow confidence intervals.

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 outbreak (Bretó et al., 2009). In order to make a precise and accurate inference, disease modelers have sought to build more realistic process models. For example, the same London measles time series from the prevaccination era has been analyzed many times – 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. 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 develops symptom (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).

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 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 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 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:

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

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 a mean 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; likewise, 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, 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.

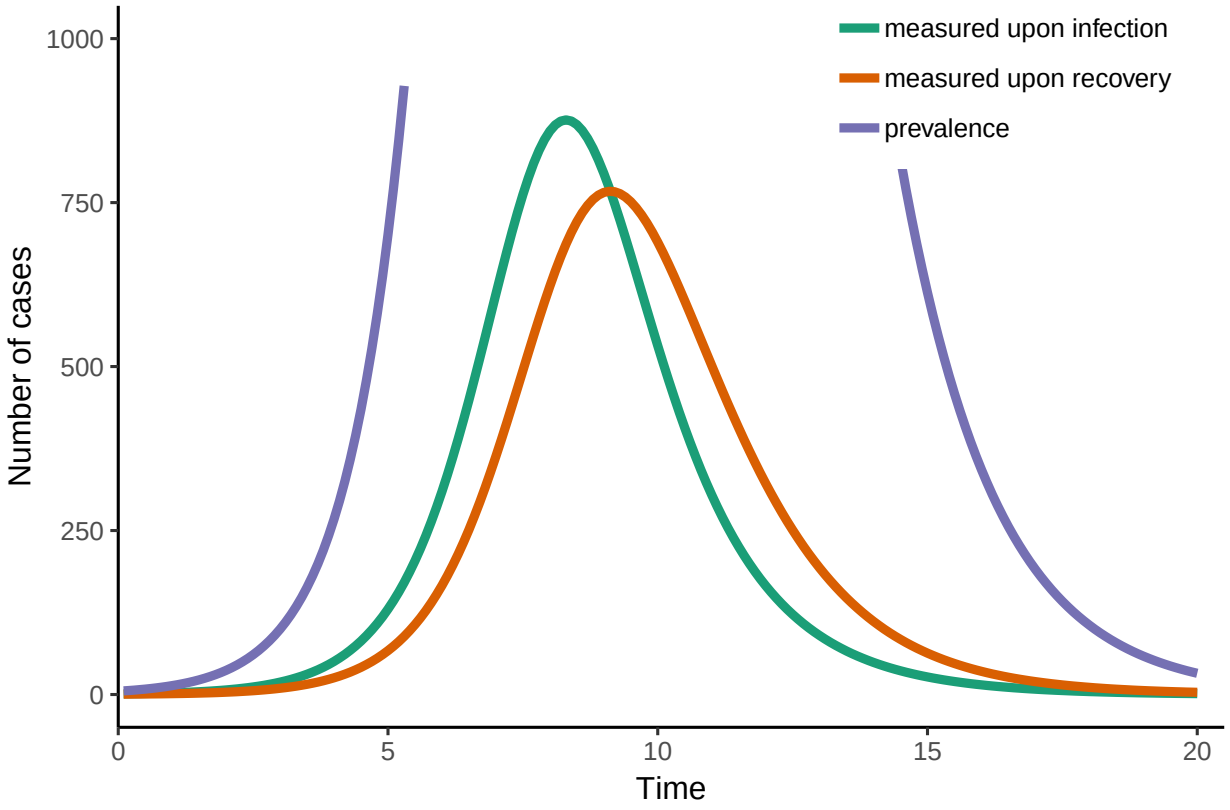

Figure 1: **A comparison of incidence measured at two different time points in infection.** A deterministic SIR model is simulated using the following parameters: $\mathcal{R}_0 = 2$, $1/\gamma = 1$ time units, $N = 1 \times 10^5$, $i0 = 1 \times 10^{-4}$, and $\Delta t = 0.1$ time units.

3 Results

First, we compare the deterministic dynamics of incidence curves, $i_1(t)$ and $i_2(t)$, in Figure 1. A lag in reporting time delays the observed epidemic peak timing and reduces 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 might not expect to be able to distinguish between the two reporting processes based on the time series alone.

Note that incidence is different from prevalence, $I(t)$, which is defined as the number of *currently* infected individuals; the observed 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 equal to the disease generation time. While they are relatively uncommon, some models do not make a clear distinction between a discrete-time observed variable and a continuous-time latent variable (Table 1).

In order to understand how assumptions about the timing of case reporting affect pa-

Study	Disease	Observed variables
Capistrán et al. (2009)	Respiratory syncytial virus	Number of hospital cases
Hooker et al. (2010)	Measles	Number of reported cases
Eisenberg et al. (2013)	Cholera	Number of reported cases
Yang et al. (2013)	Hand-foot-mouth disease	Number of reported cases
González-Parra et al. (2014)	Influenza	Number of confirmed cases

Table 1: **List of studies that fit prevalence curves to incidence data.** Incidence, defined as the number of new cases between two reporting periods, depends on the reporting time step. Prevalence, defined as the number of currently infected individuals, does not depend on the reporting time step. Some models do not make a clear distinction between these two quantities.

rameter estimates of the SIR model, we simulate observed epidemic time series (infection time series and recovery time series) 100 times and fit both infection and recovery models 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 (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, fitting the wrong model results in a clearer bias and a much lower coverage (Figure 2B); however, bias changes direction (e.g., fitting the infection model to recovery model overestimates the reproduction number instead).

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 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 we see in Figure 2B. When we allow the mean infectious period to vary, the model still over estimates the growth rate but also underestimates the mean infectious period (high γ), which decreases the overall estimate of the basic reproduction number. Likewise, 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.

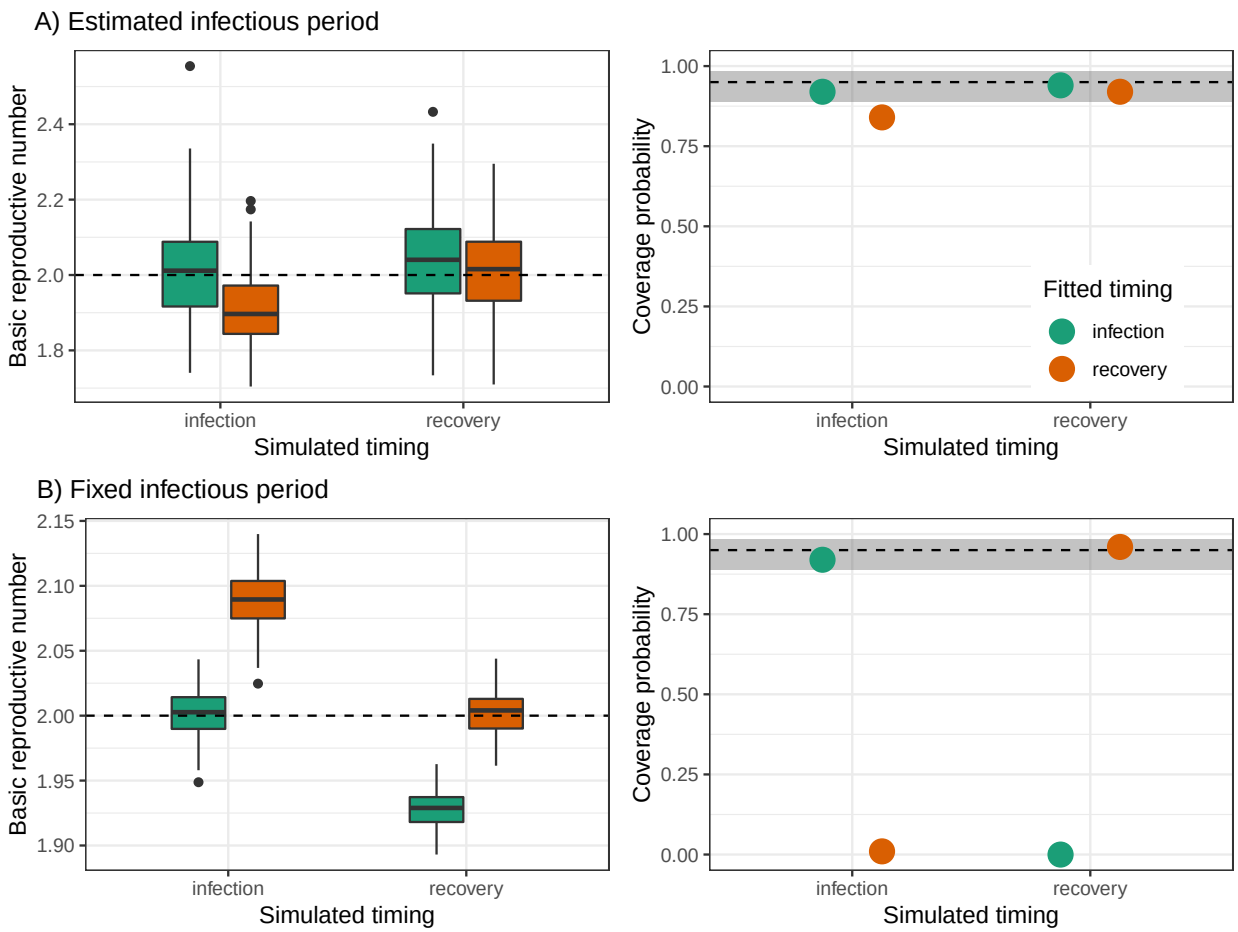

Figure 2: **A comparison of incidence measured at two different time points in infection.** We simulate infection time series and recovery time series 100 times using the following parameters: $\mathcal{R}_0 = 2$, $1/\gamma = 1$ time units, $N = 1 \times 10^5$, $i0 = 1 \times 10^{-4}$, $\rho = 0.5$, $\theta = 10$, and $\Delta t = 0.1$ time units. For each simulation, we fit the infection model and the recovery model by (A) allowing for the mean infectious period to be estimated 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 ($\mathcal{R}_0$).

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, many epidemic models make naive assumptions about how cases are reported. Here, we used a 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 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 important in mortality reporting. In particular, 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 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).

There is some empirical evidence that delays in case notification alters the shape of an epidemic curve and affects our understanding of disease transmission: (1) Postal delays in measles notification during holidays lead to sudden epidemic peaks after holidays (Fine and Clarkson, 1982); (2) Delays in reports of the foot-and-mouth infection affects the effectiveness of slaughter policies on epidemic intervention; and (3) Date of wills written and date of wills probated during the plague epidemics in London give completely different looking epidemic curves due to extremely long delays (mean of 2.9 years) between the two (Bushby, 2019). Ignoring such delays in reporting may affect conclusions about the underlying transmission mechanisms.

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 even give better coverage probabilities when wrong observation models are used. Nonetheless,

misspecifying the observation model can still affect conclusions from stochastic models and introduce bias.

Our study shows that a seemingly negligible change in the assumptions of an epidemic model can affect 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

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.

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.

Bushby, A. (2019). Demographic patterns in medieval london inferred from wills probated in the court of husting, 1259–1688. Master's thesis, McMaster University.

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.

Eisenberg, M. C., S. L. Robertson, and J. H. Tien (2013). Identifiability and estimation of multiple transmission pathways in cholera and waterborne disease. *Journal of theoretical biology 324*, 84–102.

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.

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 london. Series A, Containing papers of a mathematical and physical character 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.

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.

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.

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.