Estimating the basic reproduction number at the beginning of an outbreak under incomplete data

Sawitree Boonpatcharanon[1]
Jane Heffernan[2,3]
Hanna Jankowski[2,3]

[1] Statistics, Chulalongkorn Business School, Chulalongkorn University, Thailand
[2] Mathematics & Statistics, York University, Toronto, Canada
[3] Centre for Disease Modelling, York University, Toronto, Canada

July 14, 2021

Abstract

We compare different methods of estimating the basic reproduction number, R_0, focusing on the early stages of an epidemic, and considering weekly reports of new infecteds. We study three standard epidemiological models: SIR, SEIR, and SEAIR and examine the sensitivity of the estimators to the model structure. As some methods are developed assuming specific epidemiological models, our work adds a study of their performance in both the well- and miss-specified settings. We focus on parameters matching various types of respiratory viruses, although the general approach is easily extendable to other scenarios.

1 Introduction

The basic reproduction number, R_0, (also called the basic reproductive ratio) is defined as the expected number of new infections produced by a single (typical) infectious individual, when introduced into a totally susceptible population. R_0 is used in epidemiological studies of infectious diseases to gauge how contagious/transmissible an infectious disease is: if $R_0 < 1$, the disease will die out, and if $R_0 > 1$ infection can increase in the population. It is also used to determine how effective vaccination or other disease mitigation strategies need to be in order to protect populations from infection.

At the outset of an infectious disease outbreak, an immediate goal is to determine R_0, so that public health and healthcare decision makers can be informed. At the debut of the COVID-19 pandemic, reports of R_0 estimates were plentiful (for examples, see Zhao et al. (2020); Tuite and Fisman (2020); Knight and Mishra (2020); Mellan et al. (2020); Hilton and Keeling (2020); Price et al. (2020)). In the recent MERS-COV, 2009 H1N1, and 2003 SARS epidemics, there were also numerous studies of R_0 globally (see Nishiura et al. (2010); Chowell et al. (2011); Tuite et al. (2010); Paine et al. (2010); Fraser et al. (2009); Pourbohloul

et al. (2009); Chowell et al. (2014); Hsieh (2015); Cauchemez et al. (2014); Riley et al. (2003); Anderson et al. (2004); Wang and Ruan (2004); Dye and Gay (2003) for a small snapshot).

There are many statistical and mathematical methods that can be used to estimate R_0 (Heffernan et al., 2005; Diekmann et al., 1990, 2010; van den Driessche and Watmough, 2002; Vegvari et al., 2021; Heesterbeek and Dietz, 1996; Blumberg and Lloyd-Smith, 2013; Gallagher et al., 2020; Farrington et al., 2001; White et al., 2021). Typical studies of R_0 will thus employ a comparison of several estimators to provide increased certainty in R_0 values. Different estimators also, however, can be constructed on different assumptions related to disease characteristics i.e., serial interval, infectious period, and thus may ignore the effects of different stages of infection. For example, many R_0 estimators have been constructed to work within a Susceptible-Infectious-Recovered (SIR) disease modelling framework. Infectious diseases, however, can include periods of infection that are not infectious. The infectious period can also be split into various stages of asymptomatic and symptomatic infection, which ultimately affect the case reporting rate to public health. Therefore, methods that are based on the SIR modelling framework can project erroneous estimates of R_0, and differences in R_0 estimates may simply reflect poor estimator structure or application to data that has been misspecified. These aspects make it difficult to compare R_0 estimates to gain increased certainty.

A recent study by Gallagher et al. (2020) has discussed several nuances of different estimator methods that can affect R_0 estimates. The effect of data misspecification is only touched on briefly. Herewithin, we provide a detailed analysis of misspecified data using six different R_0 estimators. The current study is organized as follows. We first provide an introduction to three compartmental infectious disease models that we use to generate data. Six R_0 estimators are then introduced, including a discussion of their underlying compartmental model structure assumptions. We then apply each estimator to data generated from the three compartmental models. We employ parameter values representative of respiratory virus epidemics, and in particular, influenza Cowling et al. (2009); Vink et al. (2014); Park and Ryu (2018). We note that while daily data may be sometimes available during an infectious disease outbreak, it may not be complete and can include a reporting delay. We thus have chosen to use weekly case reports. Weekly case report data is also typical to outbreaks of influenza, a respiratory virus, and our chosen pathogen of study.

Table 1: Contact rate notation in epidemiological models

	parameter			
model	β	σ	ρ	γ
SIR	$S \to I : \beta I(t)/N$			$I \to R : \gamma$
SEIR	$S \to E : \beta I(t)/N$	$E \to I : \sigma$		$I \to R : \gamma$
SEAIR	$S \to E : \beta I(t)/N$	$E \to A : \sigma$	$A \to I : \rho$	$I \to R : \gamma$

Table 2: Summary of epidemiological model properties. To obtain the serial distribution, let Y_1, Y_2, Y_3 be independent exponential random variables with mean one.

model	θ	$R_0 = R_0(\theta)$	serial distribution
SIR	(β, γ)	β/γ	Y_1/γ
SEIR	(β, γ, σ)	β/γ	$Y_1/\gamma + Y_2/\sigma$
SEAIR	$(\beta, \gamma, \sigma, \rho)$	$\beta/\gamma + \beta/\rho$	$Y_1/\gamma + Y_2/\sigma + Y_3/\rho$

2 Methods

2.1 Epidemiological models

We focus on three compartmental epidemiological models, the Susceptible-Infectious-Recovered (SIR), Susceptible-Exposed-Infectious-Recovered (SEIR), and Susceptible-Exposed-Asymptomatic Infectious-Symptomatic Infectious-Recovered (SEAIR) models (model equations are provided in the Appendix A.1). All models are considered without inclusion of demography, i.e. birth and death. The total population is denoted by N with initial values of S_0 and I_0 for S and I populations, respectively, such that N is approximately equal to S_0. That is, for the SIR model, for all $t \geq 0$ it holds that $S(t) + I(t) + R(t) = S_0 + I_0 = N$. Similarly, $S(t) + E(t) + I(t) + R(t) = N$ for the SEIR model and for the SEAIR model, $S(t) + E(t) + A(t) + I(t) + R(t) = N$. We use the notation $\theta = (\beta, \sigma, \rho, \gamma)$ to denote the vector of parameters for the models, see Table 1. For each model, the associated formulas for R_0 and the serial distribution are listed in Table 2.

Data is generated using the SIR, SEIR, and SEAIR compartmental model structures using a stochastic agent-based modelling framework implemented in C++. The simulations progress at the level of individual hosts in the applicable model disease status compartments. The simulation moves forward using "event times" that are assigned to each infected individual in the population, whereby the event times correspond to the disease model compartment of which they are currently a member. Such event times correspond to infection events, when an infected individual transmits the infection to a susceptible, and times at which infected individuals progress to the next stage of infection or recover. The C++ model is based on previous work Heffernan and Wahl (2005, 2006). 1000 simulations are conducted for each of the SIR, SEIR, and SEAIR frameworks with parameters $(\beta, \sigma, \rho, \gamma) = (5/9, 1, 1, 1/3)$, giving R_0 values of $5/3$ for the SIR and SEIR models, and $20/9$ for the SEAIR model (see Table 2 for formulas). For each epidemic, the population size N is set to $10,001$ where $S(0) = 10,000$ and $I_(0) = 1$.

Figure 1 plots the number of individuals in compartment I for each model structure. The grey lines plot the simulation outcomes while the black lines plot the mean of the simulation data. Although the complete epidemic path is simulated, we assume that only the weekly number of infectious people is actually available. The epidemics are followed for 15 weeks, which covers the first 100 days of an outbreak. Simulation data is recorded at every event time. Weekly data is extracted from each simulation and saved in a data file for use for all of the R_0 estimators employed here. The blue vertical line indicates the point of inflection,

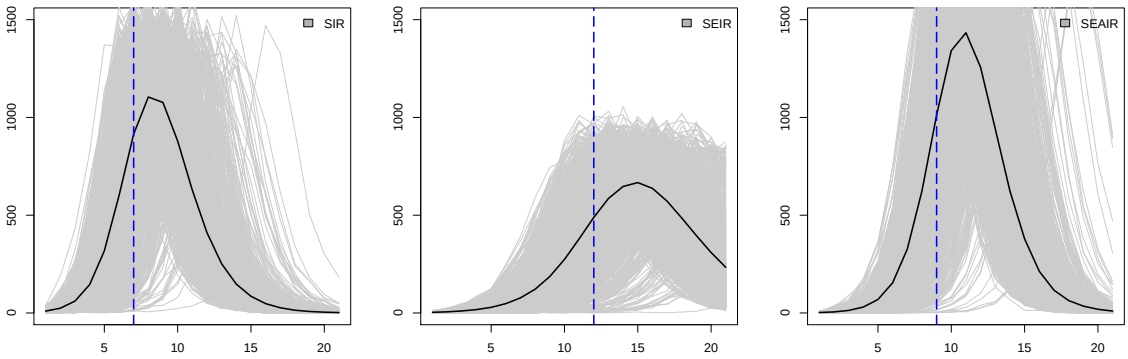

Figure 1: Plots of numbers of infectious individuals (y-axis) at time t in weeks (x-axis); from left to right: SIR, SEIR, and SEAIR. Individual simulated outbreaks from 1000 simulations are shown as grey lines, and their average is denoted as a black line. The blue vertical dashed lines show the inflection points for each model.

Table 3: Summary of estimation methods

method	
WP	serial distribution can be assumed known or can be estimated using MLE; method developed under branching process model; simple method which yields real-time estimates
secB	serial distribution assumed known (only the mean is used); method developed assuming SIR model; simple method which yields real-time estimates
ID/IDEA	serial distribution assumed known (only the mean is used); method developed assuming SIR model; simple method which yields real-time estimates
plug-and-play	serial distribution assumed unknown; method selects one of SIR/SEIR/SEAIR model; implementations available though not real-time (depending on input selection)
fullBayes	serial distribution assumed unknown; method selects one of SIR/SEIR/SEAIR model; not real-time

where the concavity of the black line changes. The inflection points were observed at 7, 12, and 9 weeks respectively for the SIR, SEIR, and SEAIR models. These points are used to determine appropriate time intervals for R_0 estimation for each model since R_0 estimates are associated with early exponential growth and can be affected by decreases in the growth rate as the epidemic continues towards and past the point of infection.

2.2 Estimating R_0

Many methods exist to estimate R_0, and we refer to White et al. (2021) for a recent review. If the transition rates in the models of Section 2.1 are known, then R_0 can be easily calculated using the formulas listed in Table 2. However, full transition rates are generally not known in practice, and hence statistical estimation methods are required. The main difficulty in estimation is that complete epidemic data is unavailable. Here, we consider six different methods of estimating R_0. These are detailed below. A summary of the methods and their key properties is also given in Table 3.

The first four (WP, secB, ID, and IDEA) are real-time methods based on simplifications of the full ODE epidemiological models. This simplification is necessitated by the fact that the full data is unobservable. In these methods, estimation of R_0 is coupled with either estimation or prior knowledge of the serial distribution. The serial distribution is the distribution of the random amount of time that an individual is infected. For example, in the SIR model, the serial distribution is exponential with mean $1/\gamma$ (see Table 2 for other models). In the literature, the serial distribution may also be referred to as the serial interval, although this most often refers to the mean of the serial distribution, or alternatively, a range indicating highly likely values from the serial distribution. As our focus here is the seasonal flu, it may be reasonable to assume that the serial distribution is known apriori. For other situations, such as new emerging diseases, such assumptions are less valid.

The two latter methods (plug-n-play and fullBayes) do not simplify the full epidemic models, but handle the issue of unobservable data by Monte Carlo simulation (plug-n-play method) or Bayesian priors with MCMC used to handle estimation due to model complexity (fullBayes method). As such, these methods are more computationally intensive. These two methods estimate the unknown transition rate parameter vector θ in the epidemic model. They do not require any prior knowledge, including prior knowledge of the serial distribution. Indeed, since the methods result in estimates of θ, these can then in turn be used to derive an estimate of the serial distribution. Furthermore, the methods assume prior knowledge of the epidemic model, in the sense that the user can decide whether the SIR, SEIR, or the SEAIR model is more appropriate for the particular disease. In contrast, the first four methods all rely on simplifications, and are not able to allow for such tailoring.

Although the last two methods are more computationally intensive and not considered "real-time", we note that modern day access to computational power is blurring this line of distinction. Our implementations of fullBayes and plug-n-play were done on a non-specialized desktop computer and without special consideration to computing time in the implementations. The time required to obtain the estimates was less than two minutes in both cases, and we do not consider this to be prohibitive. Furthermore, more careful programming could yield even faster estimates. A more detailed discussion is available in Section 3.1.

2.2.1 Maximum likelihood estimation of a branching model (WP method)

White and Pagano (2008) developed a straightforward estimation method whereby either the distribution of the serial interval is known, or, the distribution of the serial interval is estimated along with R_0. The method assumes that only the number of infectious individuals at discrete time points (e.g. daily or weekly) is observable. Both methods rely on maximum likelihood. Using our notation, and assuming that the times $t_0 = 0, t_1, t_2, \ldots, t_k$ are integers

which count, for example, the number of days or weeks, White and Pagano (2008) obtain the log-likelihood

$$\ell(R_0, p) \quad = \quad -\sum_{i=1}^{k} \mu(t_i) + \sum_{i=1}^{k} I(t_i) \log \mu(t_i),$$

where $\mu(t) = R_0 \sum_{j=1}^{\min(\kappa,t)} I(t-t_j)p(t_j)$ and p is a vector denoting the distribution of the serial interval on $t_1, \ldots, t_\kappa$. If p is assumed known (notably, this includes knowing the value of t_κ which describes the support of p) then the maximum likelihood estimate of R_0 is straightforward to compute. For example, in the SIR model, $p(t_j) = P(t_{j-1} < Y \leq t_j)/P(Y \leq t_\kappa)$ where Y is an exponential random variable with mean $1/\gamma$. If p is unknown, then White and Pagano (2008) recommend assuming a parametric distribution to simplify estimation.

The WP method assumes an underlying branching process, which is neither of the SIR/-SEIR/SEAIR models from which our data sets are generated. This model assumes, in particular, that throughout the population size "available" to be infected remains constant, which does not hold for our simulated ODE models. As such, estimates should only really be considered early on in the epidemic. In our simulations presented below, we highlight the inflection point of each epidemic, and the WP method should only really be considered valid before this time.

The method has been implemented in Obadia and Boëlle (2015), see also Obadia et al. (2017). In our simulations, we found this implementation to have some numerical instability issues, which is most likely caused by the particular parameters of our simulated data sets. This instability was particularly profound when p was assumed unknown, and most often the algorithm would not yield a solution. For this reason, we programmed our own implementation, for which we used a simple grid search. The built-in alternative optimization function in R uses the bisection method, and was very sensitive to the starting value (a small change in the starting value could change the R_0 estimate by orders of a thousand). In comparison, the grid search approach performed better, although it was still not ideal. The likelihood surface is very flat, which resulted in a non-unique MLE (we report only a default value). This property of the likelihood surface is most likely what also causes the issues in the implementation of Obadia et al. (2017).

Although not reported with our main results (which use the gamma assumption suggested by White and Pagano (2008)), we also tried other approaches to estimating the SD distribution. The easiest to implement (via a one-dimensional grid search) is the assumption that the serial distribution is any distribution with a support of at most two weeks. In Figure 2 we compare this assumption with the assumption that the serial distribution is gamma but still unknown. We also compare this with two cases where the serial distribution is known. Here, the serial distribution is exponential with either a mean of five days or a mean of three days. The true serial distribution under the SIR model is exponential with a mean of five days in our simulations, so the second setting is misspecified. Due to the weekly nature of the data and resulting discretization on the serial distribution in the WP method, the effect of this misspecification is very mild.

Furthermore, note that the log-likelihood assumes that the serial distribution is discrete, and that this discretization matches the observed data. That is, if data is observed weekly, the serial distribution is only known *on a weekly timescale*. This discretization can affect the serial distribution considerably, particularly if the timescale is quite coarse. We also note that

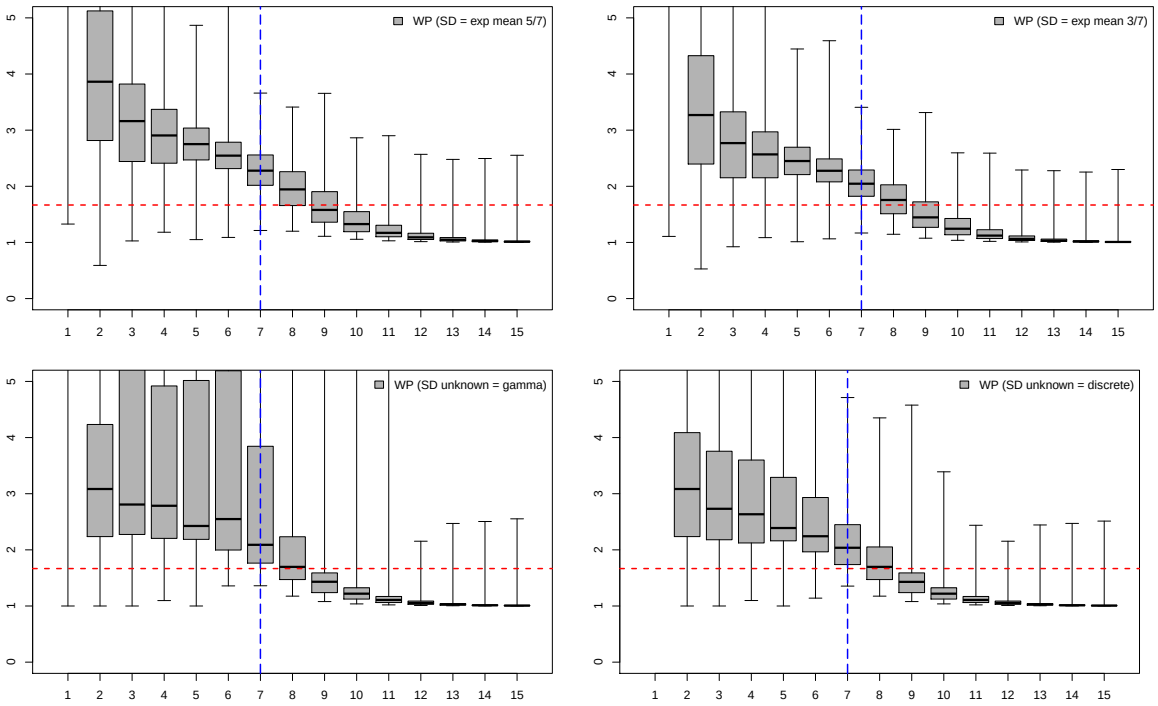

Figure 2: WP method for SIR data with four different assumptions on the serial distribution: known and correct (top left), known and incorrect (top right), unknown gamma (bottom left), unknown discrete (bottom right). The inflection point for the epidemic is marked in blue, and the true R_0 is marked as a horizontal red line.

the implementation of Obadia and Boëlle (2015) automates the discretization, and therefore we suggest that care is taken when using their built-in parametric distribution functions.

2.2.2 Sequential Bayes estimation using an SIR approximation (secB method)

Bettencourt and Riberio (2008) developed a Bayesian approach used to estimate R_0. As above, it is assumed that infectious counts are observed at periodic times such as days or weeks. The basic idea is to start with a mildly informative prior on R_0 and then update sequentially. The approach is based on the SIR model, and assumes that the mean of the serial distribution is known (under the SIR model, this is equivalent to knowing the parameter γ which is the inverse of the mean of the serial distribution). Bettencourt and Riberio (2008) note that under the SIR model

$$\begin{aligned} I(t_{j+1}) &= I(t_j)\exp\left[\gamma\int_{t_j}^{t_{j+1}}\left(R_0\frac{S(s)}{N}-1\right)ds\right] \\ &\approx I(t_j)\exp\left[(t_{j+1}-t_j)\gamma(R_t-1)\right], \end{aligned}$$

where $R_t = R_0 S(t)/N \approx R_0$ at the beginning of an infection. Using this result, seqB assumes that the conditional distribution of $I(t_{j+1})|I(t_j), R_0$ is Poisson with mean $\lambda = I(t_j)\exp\{(t_{j+1}-t_j)\gamma(R_0-1)\}$. In the approach, γ is known, and a prior is placed on R_0.

With N_0 also assumed known, posterior estimates are found using a hierarchical or sequential Bayes approach. Note that the method cannot handle data sets where there are no new infections observed in some time interval $t_{j+1} - t_j$ (as this results in a Poisson mean of zero). Therefore, the times at which infectious counts are observed must be sufficiently coarse so that all counts are non-zero (e.g. weeks instead of days). The method would also be inappropriate for situations where long intervals between cases are observed in the initial stages of the epidemic. This was observed, for example, in Canada for the first cases of Covid19.

Although the above development is based on the SIR model, the resulting approximation behaves similarly to a branching process, much like the WP method. We therefore again consider this estimator valid only in the early stages, which for our simulations translates to times prior to the inflection points of the epidemic.

The posterior distribution of R_0 will have the same support as the prior, and placing a discretized prior on R_0 makes computations relatively straightforward, since the normalizing constant of the posterior is easy to implement. In the R implementation in Obadia et al. (2017), the initial prior on R_0 is assumed to be uninformative. Their package focuses on the posterior mode, and much like their implementation of the WP method, uses a discretized version of the serial distribution (which could affect the input value of γ). We again chose to use our own implementation, and report the posterior mean which minimizes the Bayes' risk.

2.2.3 Least square estimation using incidence decay approximations (ID and IDEA methods)

Fisman et al. (2013) introduced two simplified models describing the relationship between R_0 and other epidemic parameters in the SIR model. The first of these is the incidence decay (ID) model where

$$\widetilde{I}(s) \quad = \quad R_0^s. \tag{1}$$

In the model, time is measured in units re-scaled based on the serial distribution. Recall that under the SIR model the serial distribution is exponential with mean $1/\gamma$. We then have the relationship in (1) that $\widetilde{I}(s) = I(\gamma s)$. As (1) is only valid for a short (and unknown) period of time, Fisman et al. (2013) proposed a second alternative formulation, where a decay factor was introduced in order to reflect the often observed outbreak decline. In the incidence decay and exponential adjustment (IDEA) model, the relationship becomes instead

$$\widetilde{I}(s) \quad = \quad \left(\frac{R_0}{(1+d)^s}\right)^s. \tag{2}$$

Under the ID model, we can solve (1) to obtain

$$R_0 \quad = \quad \widetilde{I}(s)^{1/s}.$$

Of course, this relationship is not valid for real data across all values of s as $\widetilde{I}(s)$ is stochastic. To obtain an estimate of R_0 least squares is an obvious option, and hence the ID estimator is the minimizer of

$$\sum_{j=1}^{k} \left(\log R_0 - \frac{1}{s_j} \log \widetilde{I}(s_j)\right)^2,$$

which yields

$$\exp\left\{\frac{1}{k}\sum_{j=1}^{k}\frac{1}{s_j}\log\widetilde{I}(s_j)\right\}. \tag{3}$$

As noted above, the number of infectious people decreases rapidly at the beginning of an outbreak, so a method based on (1) is expected to underestimate R_0. The IDEA model was introduced to overcome this issue. As in the ID model, we solve (2)

$$R_0 \;=\; \widetilde{I}(s)^{1/s}(1+d)^s,$$

and use least squares estimation to obtain its estimate. The IDEA estimator is defined then as the minimizer of

$$\sum_{j=1}^{k}\left(\log R_0 - \frac{1}{s_j}\log\widetilde{I}(s_j) - s_j\log(1+d)\right)^2.$$

Unlike in the ID model, we also need to obtain a minimizer of d to solve the optimization problem, and hence we require $k \geq 2$. Minimizing, we obtain

$$\exp\left(\frac{\sum_{j=1}^{k}s_j^2\sum_{j=1}^{k}\frac{1}{s_j}\log\widetilde{I}(s_j) - \sum_{j=1}^{k}s_j\sum_{j=1}^{k}\log\widetilde{I}(s_j)}{k\sum_{j=1}^{k}s_j^2 - (\sum_{j=1}^{k}s_j)^2}\right). \tag{4}$$

Details of these calculations are given in the Appendix. Note that the formula is not valid for $k = 1$.

Both the ID and IDEA methods are straightforward and estimate R_0 directly, as long as the mean of the serial distribution is known. The model was built under the SIR assumption, however. In our simulations we examine the effect of misspecification of the underlying epidemic model.

2.2.4 Maximum likelihood using sequential Monte Carlo for partially observed epidemics (plug-n-play method)

Maximum likelihood is one of the more popular approaches used to estimate unknown parameters in a statistical model. The general idea is to find the parameter set θ which maximizes the likelihood (probability model) evaluated at the observed data. The difficulty for our setting is that our epidemiological models (Section A.1) rely on data which is unobservable. In particular, the models require that the exact times of infections are known while we observe only daily or weekly counts of infectious individuals. The WP method (White and Pagano, 2008), which also uses maximum likelihood, gets around this issue by creating a simplified model with a likelihood which relies only on observable data. Another alternative, discussed in He et al. (2010), is to maximize the full likelihood and fill in the unobservables using many Monte Carlo simulations in a way which matches the fixed observable data points. Such an approach is often referred to as "plug-n-play".

The plug-n-play inferential method of He et al. (2010) is based on likelihood inference using sequential Monte Carlo of partially observed Markov processes (POMP), also known

as hidden Markov models or state-space models. The plug-and-play terminology comes from the fact that inference is based on Monte Carlo simulations from the model and does not require explicit expressions of the transition probabilities, which can be quite complicated. The algorithm for this method has been implemented in Nguyen et al. (2016). This software package can be accessed from the comprehensive R archive network (CRAN), see King et al. (2017). As mentioned previously, the basic idea is to generate complete epidemic data in a way which matches the observed weekly infectious observations. To simplify the implementation, complete continuous-time data is not generated but rather an approximation is generated with observations of all components at a discretized time-scale Δt (selected by the user). These discretized epidemics are generated using sequential Monte Carlo methods. An estimate of θ is then obtained via maximum likelihood using iterated filtering. The implementation in Nguyen et al. (2016) allows for the selection of the model SIR, SEIR, or SEAIR. The algorithm requires initial values for the rate parameters θ, the number of 'particles' J used in the sequential Monte Carlo method, choice of the time scale Δt, while the iterated filtering requires some further choice of algorithm settings. We refer to He et al. (2010) and Nguyen et al. (2016) for additional details. The algorithm returns estimates of θ, as well as an estimate of R_0 derived via the formula

$$R_0 \quad = \quad \beta \frac{\Delta t}{1 - e^{-\Delta t\ \gamma}},$$

regardless of the epidemiological model. We refer to the estimate thus obtained as the plug-n-play estimator. R code detailing our simulations and choices of input values is given in Appendix A.2.

2.2.5 Bayesian inference for partially observed epidemics (fullBayes method)

Similarly to the plug-n-play approach of the previous section, this is a simulation approach in which the incomplete observed data is replaced with complete data via simulations. The main difference is that the complete data is generated by placing a prior on its distribution in a Bayesian inferential approach. Some examples of epidemiological inference under the Bayesian paradigm are described in O'Neill and Roberts (1999).

In order to describe the method we need first to introduce some additional notation. We do this for the SEAIR model, as all other models are simplifications of this case. Recall that we have observed infection counts $I(t_1), \ldots, I(t_k)$ at times $t_1, \ldots, t_k$. Let m denote the vector with jth element given by $m_j = \sum_{i=1}^{j} I(t_i)$. As such, m describes the entirety of the observed data. For a time interval $[0, T]$ the complete epidemic includes much more information. Let $\tau_i^E, i \geq 2$ denote the individual times of exposure. Similarly, $\tau_i^A, i \geq 2; \tau_i^I, i \geq 2, \tau_i^R, i \geq 1$ denote the individual times of transitions into the asymptomatic, infectious, and recovered states, respectively. We assume that $m_0 = 1$. We also assume that all people who are infected in week j will recover in week $j + 1$. Furthermore, we assume that the number of exposed and asymptomatic people in week j is also equal to $m_j - m_{j-1}$. We let τ denote the epidemic path which contains all of this information.

As in O'Neill and Roberts (1999), the first infection τ_1^I is treated separately as a parameter of the model. Hence a prior $\pi_I(\tau_1^I)$ is placed on this variable. An independent prior is also placed on $\theta, \pi(\theta)$, and samples from the posterior distribution $\pi(\theta, \tau_1^I, \tau | m) \propto L(\theta, \tau_1^I | \tau, m)\pi(\theta)\pi_I(\tau_1^I)$ are obtained. The marginal distribution of $\pi(\theta, \tau_1^I, \tau | m)$ is $\pi(\theta | m)$,

which is the posterior distribution of θ given the observable data, and the distribution we are interested in.

We now calculate the likelihood $L(\theta, \tau_1^I | \tau, m)$ for the SEAIR model.

$$
\begin{aligned}
& L(\tau, m | \theta, \tau_1^I) \\
&= \left\{ \prod_{i=2}^{m_k} \frac{\beta S(\tau_i^E)}{N} \left(I(\tau_i^E) + A(\tau_i^E) \right) \right\} \left\{ \prod_{i=2}^{m_k} \sigma E(\tau_i^A) \right\} \left\{ \prod_{i=2}^{m_k} \rho A(\tau_i^I) \right\} \left\{ \prod_{i=1}^{m_{k-1}} \gamma I(\tau_i^R) \right\} \\
&\quad \times \exp \left\{ - \int_{\tau_1^I}^{t_k} \left[\beta S(t) \left(I(t) + A(t) \right) / N + \sigma E(t) + \rho A(t) + \gamma I(t) \right] dt \right\}.
\end{aligned}
$$

The joint prior distribution of the unknown rate parameters θ is made up of independent gamma distributions given by $\Gamma(\alpha, k)$ with mean k/α. We assume that α is the same for the parameters $\beta, \sigma, \rho, \gamma$, while k varies and if appropriate will be denoted by $k_\beta, k_\sigma, k_\rho, k_\gamma$. In the simulations we take $\alpha = 1$ and $k_\beta = k_\sigma = 3, k_\rho = 2, k_\gamma = 5$. The prior distribution on $-\tau_1^I$ is exponential with rate one, and this is independent from the θ vector. Calculations given in the Appendix (see Section A.4) give the posterior marginal distributions for $\pi(\tau_1^I | \theta, \tau, m)$ and $\pi(\theta | \tau, m, \tau_1^I)$ all of which have gamma distribution with closed form expressions for the parameters. Some sensitivity analysis to the prior distributions was conducted in Section A.6, and changing the prior did not visibly affect the results.

The general approach we take is now described using the following steps.

1. Use Markov chain Monte Carlo (MCMC) to simulate from $\pi(\theta, \tau, \tau_1^I | m)$.
2. From Step 1, we obtain a sequence of samples $(\tau_l, \theta_l, \tau_{1,l}^I)$ for $l = 1, \ldots, b + B$ from the posterior distribution $\pi(\theta, \tau_1^I, \tau | m)$. Here, b denotes the burn-in period for the MCMC results. To obtain an estimate of θ, from the samples $l = b + 1, b + B$, one option is to simply average the values θ_l. Instead, we treat each $(\tau_l, \theta_l, \tau_{1,l}^I)$ a sample from the full posterior model, and calculate the posterior mean of $\overline{\theta}_l$, using the formulas given in the Appendix.
3. Average the posterior means $\overline{\theta}_l, l = b + 1, \ldots, b + B$ to obtain an estimate of θ.

The final reported estimate is obtained from the estimate of θ in Step 3 using the appropriate formula in Table 2. In our simulations, we take $b = 100$ and $B = 1000$, and refer to the estimator as fullBayes.

The MCMC algorithm we use is the Metropolis-within-Gibbs. Namely, there are three main components to the posterior distribution θ, τ, and τ_1^I. In the Appendix, the posterior distributions for $\pi(\theta | \tau, \tau_1^I, m)$ and $\pi(\tau_1^I | \theta, \tau) = \pi(\tau_1^I | \theta)$ are obtain in closed form. Given one observation of $(\theta_l, \tau_l, \tau_{1,l}^I)$, the algorithm generates the next observation as follows.

1. Sample $\tau_{1,l+1}^I$ from the posterior $\pi(\tau_1^I | \theta_l)$.
2. Sample θ_{l+1} from the posterior $\pi(\theta | \tau_l, \tau_{1,l+1}^I, m)$
3. Sample τ_{l+1} using a Metropolis step:
   (a) Propose a new τ: For each $i = 1, \ldots, k$

i. τ_j^E is IID uniformly distributed on $[t_{i-1}, t_i]$ for $j = m_{i-1}, \ldots, m_i$

ii. τ_j^A is IID uniformly distributed on $[t_{i-1}, t_i]$ for $j = m_{i-1}, \ldots, m_i$

iii. τ_j^I is IID uniformly distributed on $[t_{i-1}, t_i]$ for $j = m_{i-1}, \ldots, m_i$

iv. τ_j^R is IID uniformly distributed on $[t_{i-1}, t_i]$ for $j = m_{i-2}, \ldots, m_{i-1}$

(b) Accept the proposal with probability $\min\{1, \alpha\}$ where

$$\alpha = \frac{\pi(\tau|\theta_{l+1}, \tau^I_{1,l+1}, m)g(\tau|\tau_l)}{\pi(\tau_l|\theta_{l+1}, \tau^I_{1,l+1}, m)g(\tau_l|\tau)} = \frac{L(\tau|\theta_{l+1}, \tau^I_{1,l+1}, m)}{L(\tau_l|\theta_{l+1}, \tau^I_{1,l+1}, m)},$$

noting that with the proposal distribution in (a), we have that $g(\tau|\tau_l)/g(\tau_l|\tau) = 1$. Details are provided in Appendix A.4

The chain is initialized by sampling θ from its prior distribution.

3 Results

The goal of our simulations is to study accuracy of estimation in the well-specified but also in the misspecified settings, including misspecification of the model and serial distribution. For all models we therefore consider data coming from SIR, SEIR, and SEAIR settings. We then study the methods as follows

1. WP method assuming
   - SD is known and set to exponential with mean of 5 days (correct assumption under SIR model)
   - SD is unknown and estimated from a gamma distribution with unknown mean and variance (using a grid search algorithm)
2. seqB method assuming
   - SD has a mean of 5 days (i.e. $\gamma = 7/5$)
   - SD has a mean of 3 days
3. ID and IDEA methods assuming
   - SD has a mean of 5 days (i.e. the serial interval is 5/7 in the language of Fisman et al. (2013))
   - SD has a mean of 3 days
4. plug-n-play and fullBayes methods developed assuming
   - SIR
   - SEIR
   - SEAIR

In our set-up, the outbreaks were all followed for 15 weeks, and this is the timeline given in our results. This timeline is presented only as a comparison to what is happening at the earliest stages. It also, however, improves the comparison between methods. Our comments below focus only on the time period before the inflection point (denoted as a vertical blue line for all methods).

Side-by-side boxplots summarizing our simulation results are given in Figures 3-8. Recall that the seqB method cannot handle data sets where zero values are present. We have thus

removed simulations with zero values (3 SIR, 41 SEIR, and 76 SEAIR epidemics) from the 1000 samples for seqB method. They are included for the other methods. In comparing the efficacy of the methods, we look at two main factors: bias and variability.

Figures 3 and 4 study the methods assuming that the correct model assumption is used. In Figure 3 we also include the seqB method with incorrect serial distribution for reference. Considering both bias and variance, of the methods with known SD, seqB performs best when consider the SIR model (Figure 3, second row, left column). For the SEIR and SEAIR models (Figure 4), and of the methods with unknown SD for the SIR model (Figure 3), fullBayes seems best. Although plug-n-play has small bias, in our view this is overshadowed by the extremely large variability of the method. In general, these observations carry forward into our misspecification studies.

Figure 5 focuses on methods which assume SIR and known SD, and considers misspecification of the serial distribution. Note that here the mean of the serial distribution was incorrect by only two days. Notably, in comparing the left and right columns of this figure, WP does not appear very sensitive to a change to incorrect mean, while the other methods are more sensitive.

Figures 6, 7 and 8 study model misspecification. In Figure 6 the model assumed is SIR while the data is SEIR. In Figure 7 the model assumed is SIR while the data is SEAIR. Finally, in Figure 8 the model is SEIR while the data is SEAIR. Note that only plug-n-play and fullBayes can assume the SEIR model. In Figures 6 and 7, WP with known SD performs well. With unknown SD, the estimate quality decreases in both bias and variability. In all cases fullBayes when the serial distribution is unknown performs very well.

Based on the simulations, our recommendations are as follows. In general, we recommend the fullBayes method assuming SIR, unless another model (e.g. SEIR, SEAIR) is preferred. Although the method is not considered to be real-time, we feel the computational burden is not overly onerous as estimates will be given in 1-2 minutes, depending on the computer's power and speed. In particular, fullBayes performs well even under some misspecification of the empidemiological model, and does not require prior knowledge of the serial distribution. The alternative option here is the WP method, which did not perform as well under the gamma assumption for the serial distribution in our simulation study. Alternative assumptions (such as the discrete distribution assumed in Figure 2) may also be considered, and numerical stability should be checked before applying these algorithms. If the serial distribution is known and the epidemiological model is SIR, we recommend the seqB method, but emphasize the need for sensitivity analysis.

Practitioners, however, should consider their own preferences as to bias and variability of the estimators. We note here that as this study is focused on data observed weekly, our results may not be applicable to data observed, for example, daily, as the effect of the serial distribution on the results may be different.

3.1 Computational time

Computational time is a crucial factor as real-time estimates are desirable. Table 4 shows computational time for the SEIR model for a single data set and using a 1.60GHz/8GB RAM desktop PC. The results in this work are based on fullBayes with 1000 iterations and plug-n-play with 1000 particles and 10 IF iterations, where IF stands for the iterated filtering algorithm. The fullBayes method was implemented in R, and it is possible that faster

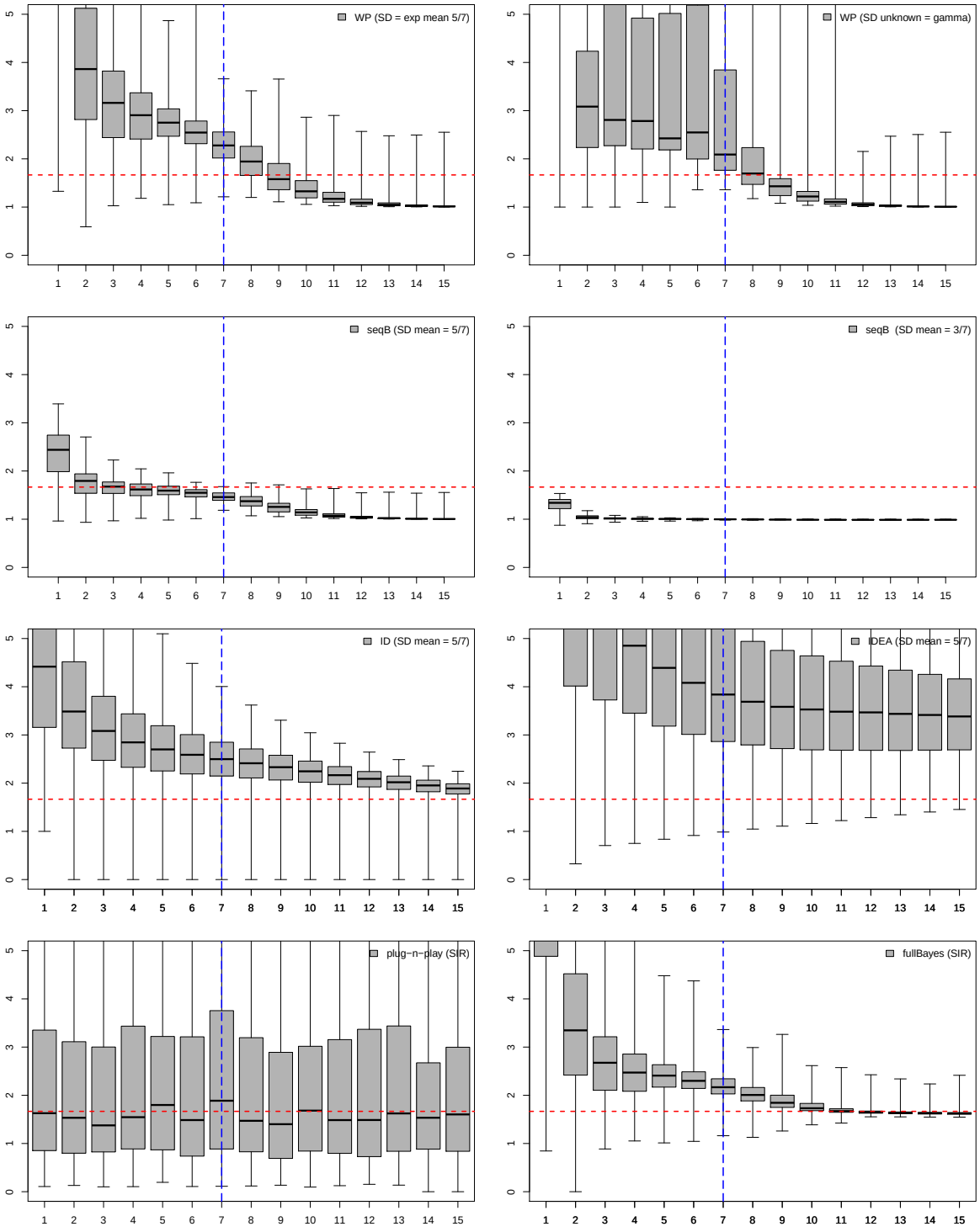

Figure 3: R_0 estimates assuming SIR model with SIR data (week on x-axis). True R_0 is given by the red dashed horizontal line, with the inflection point indicated by the blue dashed vertical line.

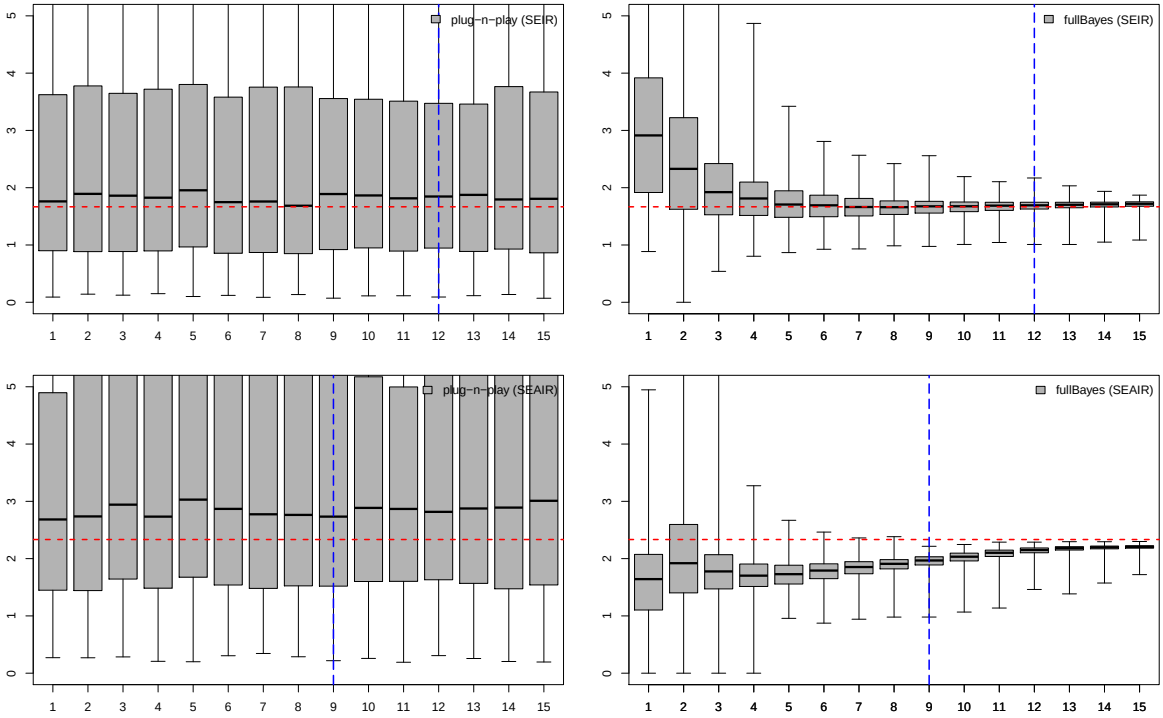

Figure 4: R_0 estimates assuming SEIR model with SEIR data in the top row and SEAIR model with SEAIR data in the bottom row (week on x-axis). True R_0 is given by the red dashed horizontal line, with the inflection point indicated by the blue dashed vertical line.

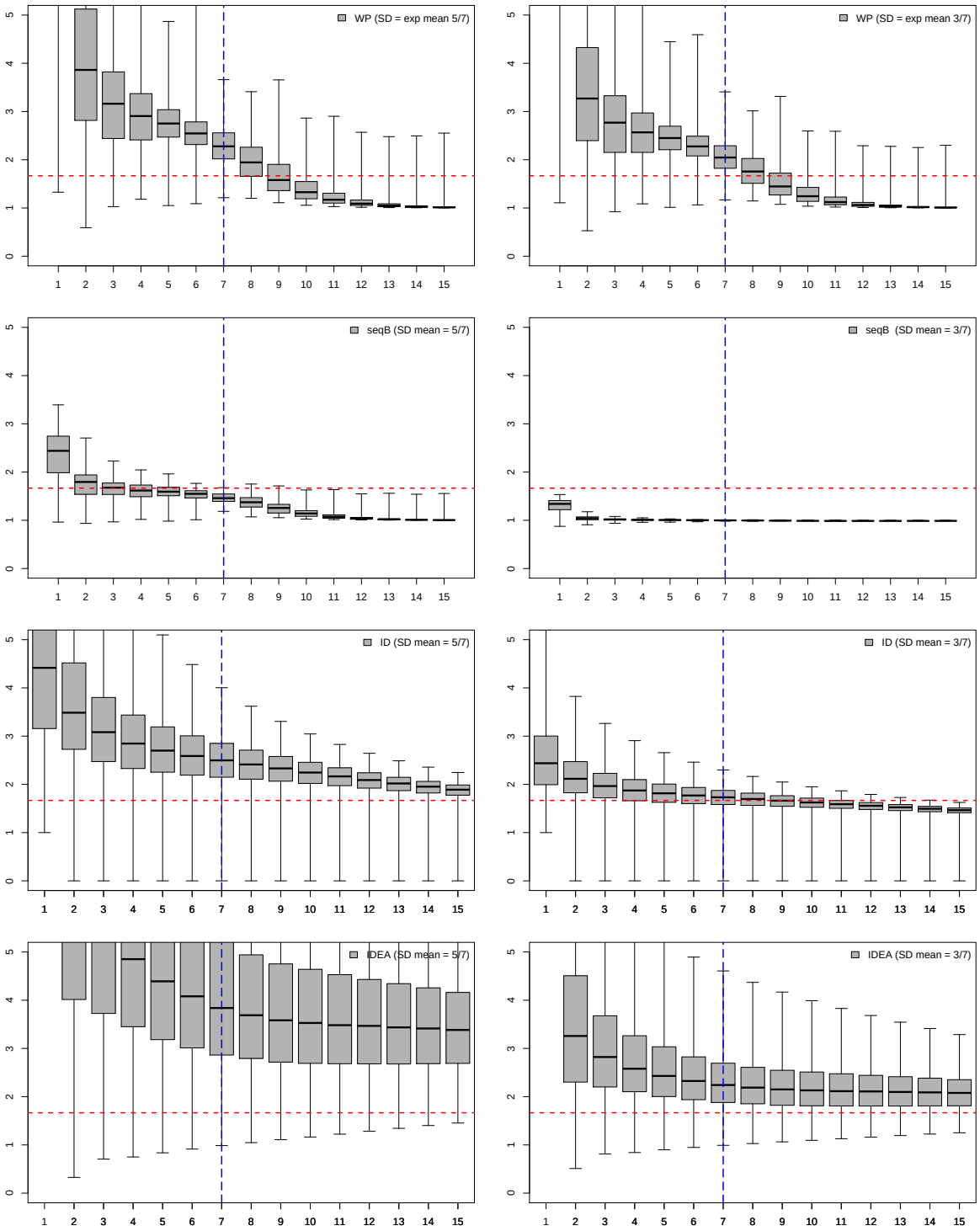

Figure 5: R_0 estimates assuming SIR model with SIR data (week on x-axis) for methods which assume SD is known with correct (left column) and incorrect (right column) assumptions on SD. True R_0 is given by the red dashed horizontal line, with the inflection point indicated by the blue dashed vertical line.

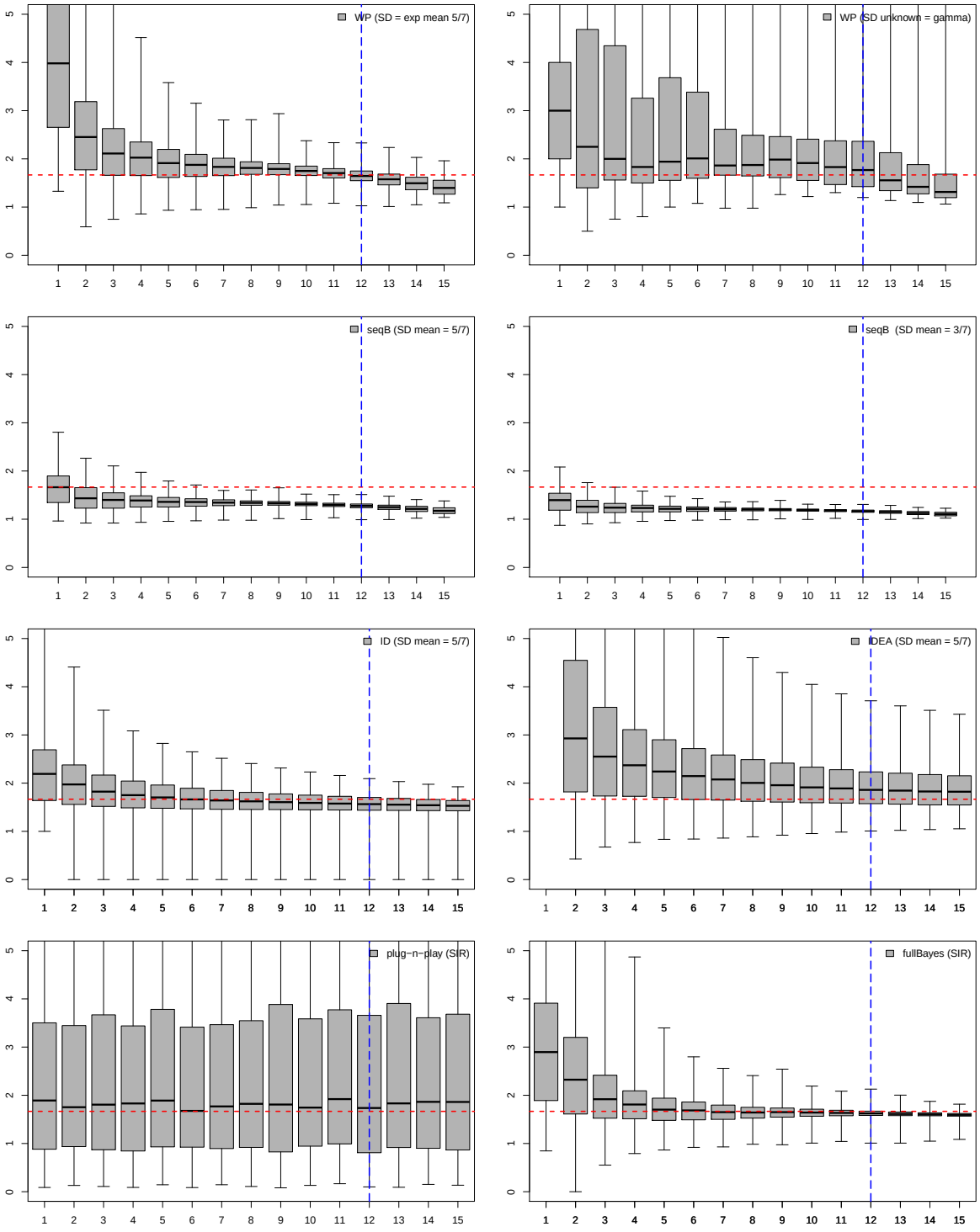

Figure 6: R_0 estimates assuming SIR model with SEIR data (week on x-axis). True R_0 is given by the red dashed horizontal line, with the inflection point indicated by the blue dashed vertical line.

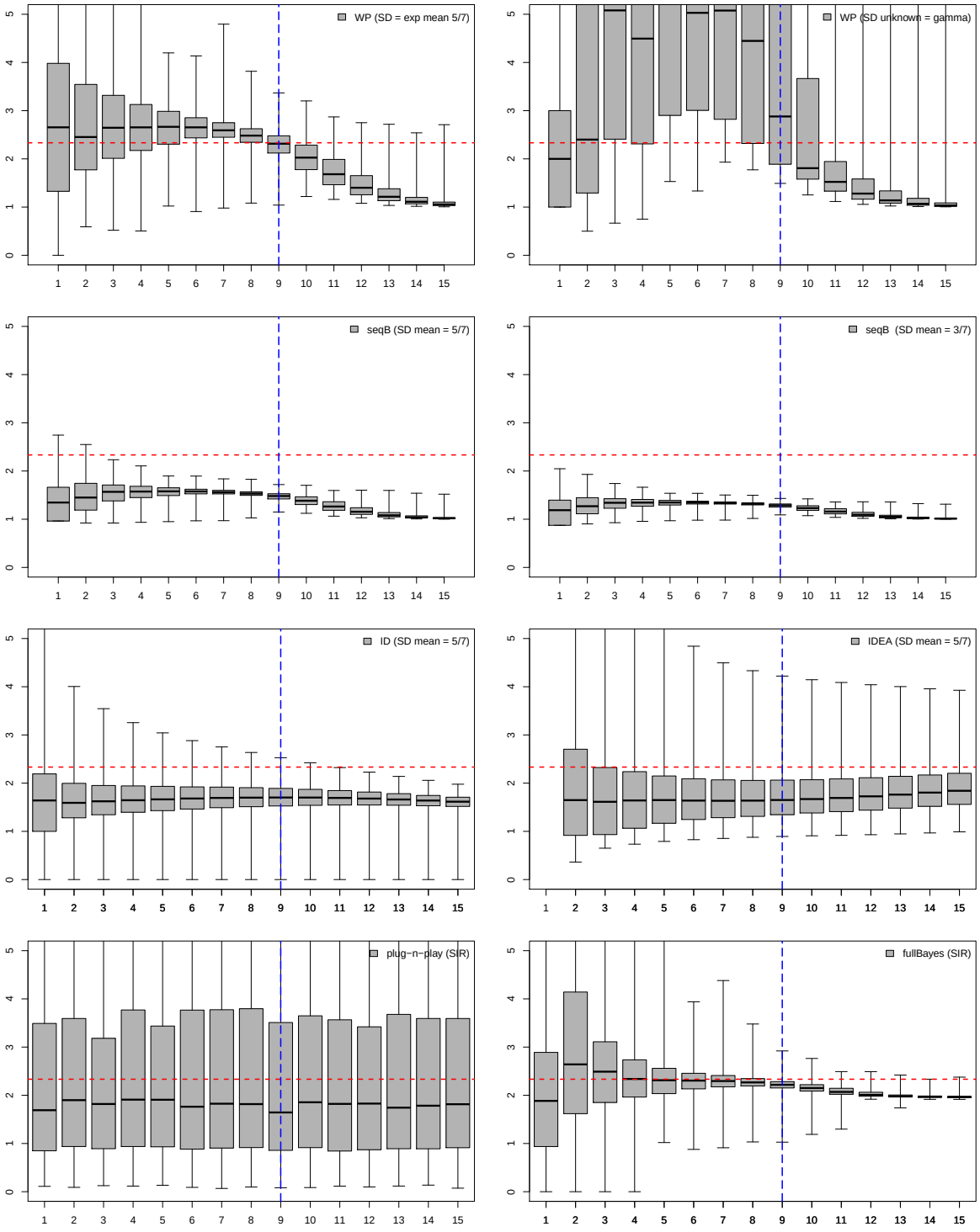

Figure 7: R_0 estimates assuming SIR model with SEAIR data (week on x-axis). True R_0 is given by the red dashed horizontal line, with the inflection point indicated by the blue dashed vertical line.

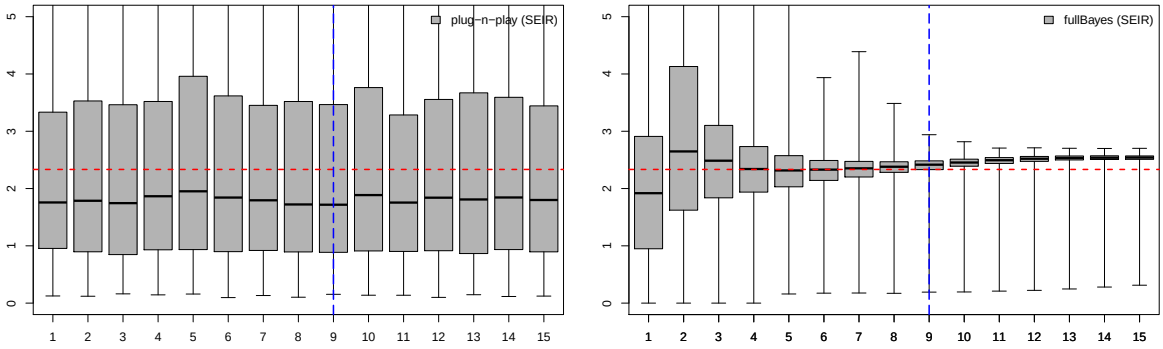

Figure 8: R_0 estimates assuming SEIR model with SEAIR data (week on x-axis). True R_0 is given by the red dashed horizontal line, with the inflection point indicated by the blue dashed vertical line.

Table 4: Computational time for the SEIR model for one data set

method		time
fullBayes	1000 iterations	1.87 minutes
	3000 iterations	3.88 minutes
	10 IF iterations	3.73 minutes
plug-n-play (1000 particles)	100 IF iterations	40 minutes
	1000 IF iterations	6 hours

implementations can be achieved using a different programming language.

4 Conclusion

The basic reproduction number, R_0, is an important parameter for estimation early in an epidemic so that public health interventions can be informed. As many estimators exist, and the assumptions of the estimators as well as their dependency on particular biological estimates i.e., the serial interval, vary between methods, it is expected that R_0 estimates will differ. It is thus important to understand what estimators provide better outcomes under both true and misspecified conditions. Since respiratory viruses (especially influenza) affect the global population every year, we have chosen to study the estimators of R_0 for these types of infections, which are typically modelled using SIR, SEIR and SEAIR compartmental models. We have also chosen to consider weekly case data, as this is characteristic of influenza and other respiratory infection outbreak reported data, globally.

We have considered six estimators that are commonly used when determining R_0 for any infectious disease outbreak. The advantages and disadvantages of each method are discussed here, including dependencies on proper estimates of the serial distribution, and the computational resources needed to run each estimator. Briefly, we find that the WP method can provide close estimates to the true R_0 value if the SD is known, but that the method

suffers numerical instability in cases when the SD is unknown for the type of data considered here; the seqB method performs well given SIR data but greatly underperforms if there is any misspecification; the ID and IDEA methods, although useful for other purposes due to their simplicity, do not outperform any of the methods studied here in terms of estimating R_0; the plug-n-play method estimates include large confidence intervals, so do not provide precise R_0 estimates; the fullBayes method is the least sensitive to model structure and misspecification. Considering both bias and variability, as well as misspecification, we find that the performance of the fullBayes estimator is best, providing estimates of R_0 that are closest to the true value under both correctly specified and misspecified cases. The approach also does not require prior knowledge of the serial distributions. We note that the choice of R_0 estimator is ultimately up to the practitioner. Our strong recommendation, however, is that if simpler methods are chosen, a full sensitivity analysis considering the misspecifications studied here, should be employed so that confidence in an R_0 estimate can be acquired.

In our analysis we have shown that some R_0 estimators can be greatly affected by even a small level of misspecification. Given that biological certainty may be lacking at the beginning of an infectious disease outbreak, the number of disease stages needed in a model and a proper distribution of the serial interval may not be known. This means that a range of R_0 results will ensue, and the accuracy of the estimates will be unclear. We therefore recommend that the fullBayes method be included in any suite of estimators used to estimate R_0 as it does not require knowledge of the serial distribution and provides close to true estimates under different model structures quickly.

Daily case reporting data has been available for the most recent COVID-19 pandemic. Daily data was not provided during the 2009 H1N1 pandemic, however. Furthermore, there may be issues with daily reporting (such as periodicity, reporting delay) whereby public health may choose to use weekly reporting data over daily data as the weekly data would be more reliable. We have thus only considered weekly case reporting data in this study as it is expected that weekly case reporting data can be expected in many future epidemics and pandemics. It is important to note that First Few Hundred (FF100) studies, whereby the first few hundred cases of a new virus are followed in detail at the beginning of an infectious disease outbreak, have been implemented during the 2009 H1N1 and COVID-19 pandemics Black et al. (2017); World Health Organization and others (2020); McLean et al. (2010); Boddington et al. (2020); van Gageldonk-Lafeber et al. (2012); England (2009); Ghani et al. (2009); Pandemic Influenza (2014). In these cases the serial distribution, and the need to consider exposed and/or asymptomatic periods of infection can be quickly determined, enabling realization of earlier and more certain estimates of R_0 early on. Given that First Few Hundred protocols are not implemented in much of the globe, weekly case report data however may still be considered the norm for future pandemics.

In our current study we have assumed perfect data with no unobserved infections, no reporting delay, and no data collection bias. These issues are intuitively expected to affect R_0 estimates. We venture to continue our study of R_0 estimation while considering these aspects in our epidemiological data sets.

A Appendix

A.1 Disease models

We employ SIR, SEIR and SEAIR model structures, where S, E, A, I and R denotes susceptible, exposed (infected, no symptoms, not infectious), asymptomatic infected (infected, no symptoms, not infectious), symptomatic infected (infected, symptomatic, infectious), and recovered individuals, respectively. The ODEs governing our models are given below.

A.1.1 SIR model

$$\begin{aligned}
\frac{dS(t)}{dt} &= -\beta S(t)\frac{I(t)}{N} \\
\frac{dI(t)}{dt} &= \beta S(t)\frac{I(t)}{N} - \gamma I(t) \\
\frac{dR(t)}{dt} &= \gamma I(t) \\
N &= S(t) + I(t) + R(t)
\end{aligned}$$

A.1.2 SEIR model

$$\begin{aligned}
\frac{dS(t)}{dt} &= -\beta S(t)\frac{I(t)}{N} \\
\frac{dE(t)}{dt} &= \beta S(t)\frac{I(t)}{N} - \alpha E(t) \\
\frac{dI(t)}{dt} &= \alpha E(t) - \gamma I(t) \\
\frac{dR(t)}{dt} &= \gamma I(t) \\
N &= S(t) + E(t) + I(t) + R(t)
\end{aligned}$$

A.1.3 SEAIR model

$$\begin{aligned}
\frac{dS(t)}{dt} &= -\beta S(t)\frac{I(t)+A(t)}{N} \\
\frac{dE(t)}{dt} &= \beta S(t)\frac{I(t)+A(t)}{N} - \alpha E(t) \\
\frac{dA(t)}{dt} &= \alpha E(t) - \rho A(t) \\
\frac{dI(t)}{dt} &= \rho A(t) - \gamma I(t) \\
\frac{dR(t)}{dt} &= \gamma I(t) \\
N &= S(t) + E(t) + A(t) + I(t) + R(t)
\end{aligned}$$

A.2 Least squares estimation for the IDEA method

From (2), our objective function is

$$Q = \sum_{j=1}^{k}\left(\log R_0 - \frac{1}{s_j}\log \widetilde{I}(s_j) - s_j \log(1+d)\right)^2.$$

Let $\eta = \log R_0$ and $\xi = \log(1+d)$ and note that both of these relationships are monotone increasing. We minimize Q by setting $\frac{\partial Q}{\partial \eta} = 0$ and $\frac{\partial Q}{\partial \xi} = 0$, obtaining two equations

$$\begin{aligned}
\xi &= \frac{k\eta - \sum_{j=1}^{k}\frac{1}{s_j}\log \widetilde{I}(s_j)}{\sum_{j=1}^{k} s_j}, \\
\xi &= \frac{\eta \sum_{j=1}^{k} s_j - \sum_{j=1}^{k}\log \widetilde{I}(s_j)}{\sum_{j=1}^{k} s_j^2}.
\end{aligned}$$

Solving for $\eta = \log R_0$ we thus find

$$\widehat{R}_{\mathsf{IDEA}} = \exp\left(\frac{\sum_{j=1}^{k} s_j^2 \sum_{j=1}^{k}\frac{1}{s_j}\log \widetilde{I}(s_j) - \sum_{j=1}^{k} s_j \sum_{j=1}^{k}\log \widetilde{I}(s_j)}{k\sum_{j=1}^{k} s_j^2 - (\sum_{j=1}^{k} s_j)^2}\right).$$

A.3 R code to implement plug-n-play estimation for SIR model

```
library(pomp)
set.seed(1)
dat <- read.csv("Weekly_NewnumInf_SIR.csv")
dat <- dat[1,]
dat <- as.numeric(dat)
dat <- cumsum(dat)
```

```
nc <- length(dat)
rho <- 1   ## all cases are reported
difft <- 1/7  ## do filtering for every day

resbeta <- rep(0,nc)
resgamma <- rep(0,nc)
R0e <- rep(0,nc)
R0 <- rep(0,nc)
  for (j in 1:nc){
    data <- c(1,as.numeric(dat[1:j]))
    date <- as.numeric(seq(0,j,1))
    infdat <- cbind(date,data)
    infdat <- as.data.frame(infdat)
    colnames(infdat) <- c("week","cases")

    ######### function ############
    sir.proc.sim <- function (S, I, R, H, beta, gamma, delta.t, ...) {
      N <- sum(S,I,R)
      foi <- beta*I/N
      trans <- c(reulermultinom(n=1,size=S,rate=foi,dt=delta.t),
                 reulermultinom(n=1,size=I,rate=gamma,dt=delta.t))
      S = S-trans[1]
      I = I+trans[1]-trans[2]
      R = R+trans[2]
      H = H+trans[2]
      c(S=S, I=I, R=R, H=H)
    }
    f <- function(t,S, I, R, beta, gamma){
      N <- sum(S, I, R)
      foi <- beta*I/N
      terms <- c(
        S*foi,
        I*gamma
      )
      terms <- unname(terms)
      c(
        S=-terms[1],
        I=terms[1]-terms[2],
        R=terms[2],
        H=terms[2]
      )
    }
    sir_dmeas <- function (cases, H, rho, log, ...) {
        #cases = sum(infdat[,2])
        dbinom(x=cases, size=H, prob=rho, log=log)
    }
```

```
    sir_rmeas <- function (H, rho, ...) {
        cases=rbinom(n=1, size=H, prob=rho)
    }
    flu.sir <- pomp(data= infdat,
                    times="week",
                    t0=0,
                    params=c(rho=1,gamma=mean(rgamma(n=20,1,5)),beta=mean(rgamma(
                             S.0=10000,R.0=0,I.0=1,H.0=0),
                    rmeasure=sir_rmeas,
                    dmeasure=sir_dmeas,
                    rprocess=euler(sir.proc.sim, delta.t=difft)
    )
    ######################################## end function ##############################

    para <- coef(flu.sir)
    simpar <- c("beta","gamma")
    fit <- mif2(flu.sir,Nmif=5,
                rw.sd=rw.sd(beta=0.01,gamma=0.005),
                cooling.fraction.50=0.01,
                Np=1000)
    suppressWarnings(fit)
    resbeta[j] <- as.numeric(coef(fit)[3])
    resgamma[j] <- as.numeric(coef(fit)[2])
    IP <- difft/(1-exp(-difft*resgamma[j]))
    R0e[j] <- resbeta[j]*IP
    R0[j]  <- resbeta[j]/resgamma[j]
}
```

A.4 Posterior distributions

$$
\begin{aligned}
&L(\tau, m|\theta, \tau_1^I) \\
&= \left\{\prod_{i=2}^{m_k} \frac{\beta S(\tau_i^E)}{N}\left(I(\tau_i^E)+A(\tau_i^E)\right)\right\}\left\{\prod_{i=2}^{m_k}\sigma E(\tau_i^A)\right\}\left\{\prod_{i=2}^{m_k}\rho A(\tau_i^I)\right\}\left\{\prod_{i=1}^{m_{k-1}}\gamma I(\tau_i^R)\right\} \\
&\qquad \times \exp\left\{-\int_{\tau_1^I}^{t_k}\left[\beta S(t)\left(I(t)+A(t)\right)/N+\sigma E(t)+\rho A(t)+\gamma I(t)\right]dt\right\}.
\end{aligned}
$$

First, we calculate the posterior distributions of each of the elements of θ. For ease of

exposition, in each calculation we drop the subscript on the scale parameter k in the priors.

$$\begin{aligned}
&\pi(\beta|\sigma,\rho,\gamma,\tau_1^I,\tau) \\
&\quad\propto\quad L(\tau,m|\theta,\tau_1^I)\pi(\beta) \\
&\quad\propto\quad \left\{\prod_{i=2}^{m_k}\frac{\beta S(\tau_i^E)}{N}\left(I(\tau_i^E)+A(\tau_i^E)\right)\right\}\exp\left\{-\int_{\tau_1^I}^{t_k}\beta S(t)\left(I(t)+A(t)\right)/N dt\right\}\beta^{\alpha-1}\exp(-k\beta) \\
&\quad\propto\quad \beta^{(\alpha+m_k)-1}\exp\left\{-\beta\left(k+\int_{\tau_1^I}^{t_k}S(t)\left(I(t)+A(t)\right)/N dt\right)\right\}.
\end{aligned}$$

Hence the posterior distribution for β is gamma with shape parameter $\alpha+m_k$ and scale parameter $k+\int_{\tau_1^I}^{t_k}S(t)\left(I(t)+A(t)\right)/N dt$.

$$\begin{aligned}
\pi(\sigma|\beta,\rho,\gamma,\tau_1^I,\tau) &\propto L(\tau,m|\theta,\tau_1^I)\pi(\sigma) \\
&\propto \left\{\prod_{i=2}^{m_k}\sigma E(\tau_i^A)\right\}\exp\left\{-\int_{\tau_1^I}^{t_k}\sigma E(t)dt\right\}\sigma^{\alpha-1}\exp(-k\sigma) \\
&\propto \sigma^{(m_k+\alpha)-1}\exp\left\{-\sigma\left(k+\int_{\tau_1^I}^{t_k}E(t)dt\right)\right\}
\end{aligned}$$

Hence the posterior distribution for σ is gamma with shape parameter $\alpha+m_k$ and scale parameter $k+\int_{\tau_1^I}^{t_k}E(t)dt$.

$$\begin{aligned}
\pi(\rho|\beta,\sigma,\gamma,\tau_1^I,\tau) &\propto L(\tau,m|\theta,\tau_1^I)\pi(\rho) \\
&\propto \left\{\prod_{i=2}^{m_k}\rho A(\tau_i^I)\right\}\exp\left\{-\int_{\tau_1^I}^{t_k}\rho A(t)dt\right\}\rho^{\alpha-1}\exp(-k\rho) \\
&= \rho^{(\alpha+m_k)-1}\exp\left\{-\rho\left(k+\int_{\tau_1^I}^{t_k}A(t)dt\right)\right\}
\end{aligned}$$

Hence the posterior distribution for σ is gamma with shape parameter $\alpha+m_k$ and scale parameter $k+\int_{\tau_1^I}^{t_k}A(t)dt$.

$$\begin{aligned}
\pi(\gamma|\beta,\rho,\gamma,\tau_1^I,\tau) &\propto L(\tau,m|\theta,\tau_1^I)\pi(\gamma) \\
&\propto \left\{\prod_{i=1}^{m_{k-1}}\gamma I(\tau_i^R)\right\}\exp\left\{-\int_{\tau_1^I}^{t_k}\gamma I(t)dt\right\}\gamma^{\alpha-1}\exp(-k\gamma) \\
&\propto \gamma^{(\alpha+m_{k-1})-1}\exp\left\{-\gamma\left(k+\int_{\tau_1^I}^{t_k}I(t)dt\right)\right\}
\end{aligned}$$

Hence the posterior distribution for γ is gamma with shape parameter $\alpha+m_k$ and scale parameter $k+\int_{\tau_1^I}^{t_k}I(t)dt$.

Lastly, we calculate the posterior distribution of $-\tau_1^I$, with a prior distribution of exponential, rate one.

$$\begin{aligned}\pi(-\tau_1^I|\theta,\tau) &\propto L(\tau,m|\theta,\tau_1^I)\pi(-\tau_1^I)\\ &\propto \exp\left(-\int_{\tau_1^I}^{\tau_2^I}\frac{\beta}{N}S(t)(I(t)+A(t))+\sigma E(t)+\rho A(t)+\gamma I(t)dt\right)\exp(\tau_1^I).\end{aligned}$$

For $\tau_1^I \le t < \tau_2^I$, we have that $S(t) = N$, $I(t) = 1$, $E(t), A(t) = 0$, and hence

$$\int_{\tau_1^I}^{\tau_2^I}\frac{\beta}{N}S(t)(I(t)+A(t))+\sigma E(t)+\rho A(t)+\gamma I(t)dt \;=\; (\beta+\gamma)(\tau_2^I-\tau_1^I),$$

from which it follows that

$$\begin{aligned}\pi(-\tau_1^I|\beta,\sigma,\rho,\gamma,\tau) &\propto \exp\left(-(\beta+\gamma)(\tau_2^I-\tau_1^I)+\tau_1^I\right)\\ &\propto \exp\left((\beta+\gamma+1)\tau_1^I\right),\end{aligned}$$

and hence the posterior of $-\tau_1^I$ is exponential with rate $\beta+\gamma+1$. Note that this formula is the same for all models.

A.5 Symmetric Proposal

We need to show that $g(\tau|\tau_l)/g(\tau_l|\tau) = 1$. To do this, we use the fact that $g(\tau|\tau_l)$ does not depend on τ_l. Moreover, $g(\tau)$ is a product of uniform distributions, so $g(\tau)$ also does not depend on τ. Therefore $g(\tau|\tau_l) = c$ for some constant c, and hence $g(\tau|\tau_l) = g(\tau_l|\tau) = 1$.

A.6 Sensitivity to Prior

As mentioned previously, the joint prior distribution of the unknown rate parameters θ is made up of independent gamma distributions given by $\Gamma(\alpha, k)$ with mean k/α. In the main text, we assume that α is the same for the parameters $\beta, \sigma, \rho, \gamma$, while k varies and if appropriate will be denoted by $k_\beta, k_\sigma, k_\rho, k_\gamma$. In the simulations we took these to be $\alpha = 1$ and $k_\beta = k_\sigma = 3, k_\rho = 2, k_\gamma = 5$. The prior distribution on $-\tau_1^I$ is exponential with rate one, and this is independent from the θ vector. In Figure 9, we compare the results in the main text with results repeating the method with a different prior distribution for the SIR/SEIR/SEAIR data assuming SIR/SEIR/SEAIR models respectively. The modified prior for the comparison is $k_\beta = 9/4, k_\gamma = 3$. These were chosen as alternative reasonable parameters for the flu. The plots show that there was very little change between the two versions.

References

Anderson, R. M., Fraser, C., Ghani, A. C., Donnelly, C. A., Riley, S., Ferguson, N. M., Leung, G. M., Lam, T. H., and Hedley, A. J. (2004). Epidemiology, transmission dynamics and control of SARS: The 2002–2003 epidemic. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 359(1447):1091–1105.

Bettencourt, L. M. and Riberio, R. M. (2008). Real time Bayesian estimation of the epidemic potential of emerging infectious diseases. *PLOS ONE*, 3.

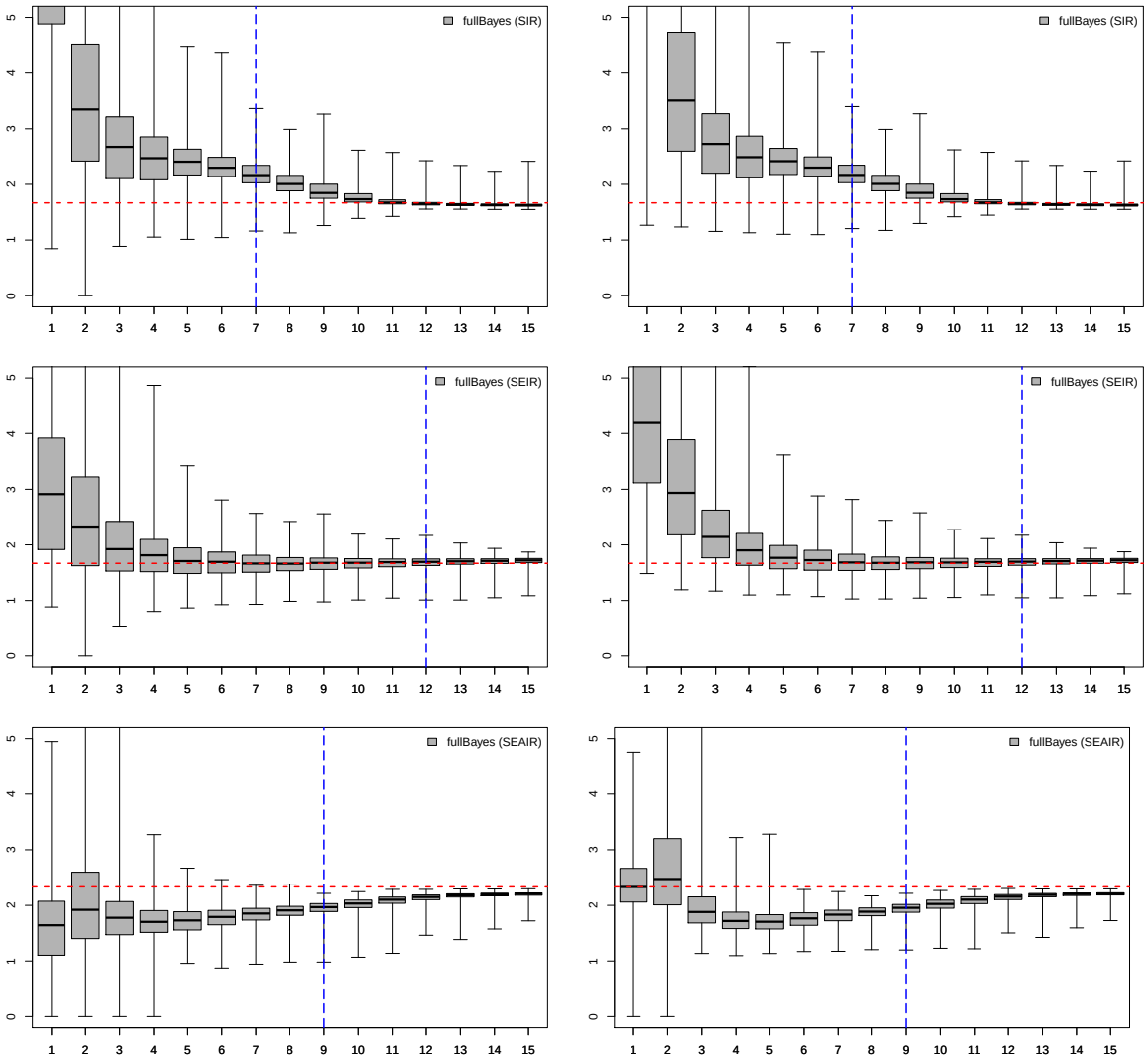

Figure 9: Comparison of the fullBayes method for SIR, SEIR, and SEAIR data with two different prior distributions: same as main text is on the right and the modified version is on the left. The inflection point for the epidemic is marked in blue, and the true R_0 for the data is marked as a horizontal red line.

Black, A. J., Geard, N., McCaw, J. M., McVernon, J., and Ross, J. V. (2017). Characterising pandemic severity and transmissibility from data collected during first few hundred studies. *Epidemics*, 19:61–73.

Blumberg, S. and Lloyd-Smith, J. O. (2013). Comparing methods for estimating R0 from the size distribution of subcritical transmission chains. *Epidemics*, 5(3):131–145.

Boddington, N. L., Charlett, A., Elgohari, S., Walker, J. L., McDonald, H. I., Byers, C., Coughlan, L., Vilaplana, T. G., Whillock, R., Sinnathamby, M., et al. (2020). COVID-19 in Great Britain: epidemiological and clinical characteristics of the first few hundred (FF100) cases: a descriptive case series and case control analysis. *MedRxiv*.

Cauchemez, S., Fraser, C., Van Kerkhove, M. D., Donnelly, C. A., Riley, S., Rambaut, A., Enouf, V., van der Werf, S., and Ferguson, N. M. (2014). Middle east respiratory syndrome coronavirus: quantification of the extent of the epidemic, surveillance biases, and transmissibility. *The Lancet Infectious Diseases*, 14(1):50–56.

Chowell, G., Blumberg, S., Simonsen, L., Miller, M. A., and Viboud, C. (2014). Synthesizing data and models for the spread of MERS-CoV, 2013: key role of index cases and hospital transmission. *Epidemics*, 9:40–51.

Chowell, G., Echevarría-Zuno, S., Viboud, C., Simonsen, L., Tamerius, J., Miller, M. A., and Borja-Aburto, V. H. (2011). Characterizing the epidemiology of the 2009 influenza A/H1N1 pandemic in mexico. *PLoS Med*, 8(5):e1000436.

Cowling, B. J., Fang, V. J., Riley, S., Peiris, J. M., and Leung, G. M. (2009). Estimation of the serial interval of influenza. *Epidemiology (Cambridge, Mass.)*, 20(3):344.

Diekmann, O., Heesterbeek, J., and Roberts, M. G. (2010). The construction of next-generation matrices for compartmental epidemic models. *Journal of the Royal Society Interface*, 7(47):873–885.

Diekmann, O., Heesterbeek, J. A. P., and Metz, J. A. (1990). On the definition and the computation of the basic reproduction ratio R0 in models for infectious diseases in heterogeneous populations. *Journal of Mathematical Biology*, 28(4):365–382.

Dye, C. and Gay, N. (2003). Modeling the sars epidemic. *Science*, 300(5627):1884–1885.

England, H. P. A. (2009). *"First Few Hundred" Project Epidemiological Protocols for Comprehensive Assessment of Early Swine Influenza Cases in the United Kingdom*.

Farrington, C. P., Kanaan, M. N., and Gay, N. J. (2001). Estimation of the basic reproduction number for infectious diseases from age-stratified serological survey data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 50(3):251–292.

Fisman, D. N., Hauck, T. S., Tuite, A. R., and Greer, A. L. (2013). An idea for short term outbreak projection: Nearcasting using the basic reproduction number. *PLOS ONE*, 8.

Fraser, C., Donnelly, C. A., Cauchemez, S., Hanage, W. P., Van Kerkhove, M. D., Hollingsworth, T. D., Griffin, J., Baggaley, R. F., Jenkins, H. E., Lyons, E. J., et al. (2009). Pandemic potential of a strain of influenza A (H1N1): early findings. *Science*, 324(5934):1557–1561.

Gallagher, S., Chang, A., and Eddy, W. F. (2020). Exploring the nuances of R0: Eight estimates and application to 2009 pandemic influenza. *arXiv preprint arXiv:2003.10442*.

Ghani, A., Baguelin, M., Griffin, J., Flasche, S., van Hoek, A. J., Cauchemez, S., Donnelly, C., Robertson, C., White, M., Truscott, J., et al. (2009). The early transmission dynamics of H1N1pdm influenza in the united kingdom. *PLoS Currents*, 1.

He, D., Ionides, E. L., and King, A. A. (2010). Plug-and-play inference for disease dynamics: measles in large and small populations as a case study. *J.R. Soc. Interface*, 7.

Heesterbeek, J. and Dietz, K. (1996). The concept of R0 in epidemic theory. *Statistica Neerlandica*, 50(1):89–110.

Heffernan, J. M., Smith, R. J., and Wahl, L. M. (2005). Perspectives on the basic reproductive ratio. *Journal of the Royal Society Interface*, 2(4):281–293.

Heffernan, J. M. and Wahl, L. M. (2005). Monte Carlo estimates of natural variation in HIV infection. *Journal of Theoretical Biology*, 236(2):137–153.

Heffernan, J. M. and Wahl, L. M. (2006). Natural variation in HIV infection: Monte Carlo estimates that include CD8 effector cells. *Journal of Theoretical Biology*, 243(2):191–204.

Hilton, J. and Keeling, M. J. (2020). Estimation of country-level basic reproductive ratios for novel coronavirus (sars-cov-2/covid-19) using synthetic contact matrices. *PLoS Computational Biology*, 16(7):e1008031.

Hohle, M. and Jorgensen, E. (2002). Estimating parameters for stochastic epidemics. *Dina Research Report*, (102).

Hsieh, Y.-H. (2015). 2015 middle east respiratory syndrome coronavirus (MERS-CoV) nosocomial outbreak in south korea: insights from modeling. *PeerJ*, 3:e1505.

King, A. A., Ionides, E. L., and Breto, C. (2017). *Statistical Inference for Partially Observed Markov Processes*. R package version 1.12.

Knight, J. and Mishra, S. (2020). Estimating effective reproduction number using generation time versus serial interval, with application to COVID-19 in the Greater Toronto Area, Canada. *Infectious Disease Modelling*, 5:889–896.

McLean, E., Pebody, R., Campbell, C., Chamberland, M., Hawkins, C., Nguyen-Van-Tam, J., Oliver, I., Smith, G., Ihekweazu, C., Bracebridge, S., et al. (2010). Pandemic (H1N1) 2009 influenza in the UK: clinical and epidemiological findings from the first few hundred (FF100) cases. *Epidemiology & Infection*, 138(11):1531–1541.

Mellan, T. A., Hoeltgebaum, H. H., Mishra, S., Whittaker, C., Schnekenberg, R. P., Gandy, A., Unwin, H. J. T., Vollmer, M. A., Coupland, H., Hawryluk, I., et al. (2020). Report 21: Estimating covid-19 cases and reproduction number in Brazil. *medRxiv*.

Nguyen, D., Ionides, E. L., and King, A. A. (2016). Statistical inference for partially observed markov processes via the R package pomp. *Journal of Statistical Software*, 69.

Nishiura, H., Chowell, G., Safan, M., and Castillo-Chavez, C. (2010). Pros and cons of estimating the reproduction number from early epidemic growth rate of influenza a (h1n1) 2009. *Theoretical Biology and Medical Modelling*, 7(1):1–13.

Obadia, T. and Boëlle, P. (2015). *R0: Estimation of R0 and Real-Time Reproduction Number from Epidemics*. R package version 1.2-6.

Obadia, T., Haneef, R., and Boëlle, P. (2017). The R0 package: a toolbox to estimate reproduction numbers for epidemic outbreaks. *BMC Med Inform Decis Mak*, 12.

O'Neill, P. D. and Roberts, G. O. (1999). Bayesian inference for partially observed stochastic epidemics. *Journal of the Royal Statistical Society Series A*, 162:121–129.

Paine, S., Mercer, G., Kelly, P., Bandaranayake, D., Baker, M., Huang, Q., Mackereth, G., Bissielo, A., Glass, K., and Hope, V. (2010). Transmissibility of 2009 pandemic influenza A (H1N1) in New Zealand: effective reproduction number and influence of age, ethnicity and importations. *Eurosurveillance*, 15(24):19591.

Pandemic Influenza (2014). Australian health management plan for pandemic influenza.

Park, J.-E. and Ryu, Y. (2018). Transmissibility and severity of influenza virus by subtype. *Infection, Genetics and Evolution*, 65:288–292.

Pourbohloul, B., Ahued, A., Davoudi, B., Meza, R., Meyers, L. A., Skowronski, D. M., Villaseñor, I., Galván, F., Cravioto, P., Earn, D. J., et al. (2009). Initial human transmission dynamics of the pandemic (H1N1) 2009 virus in North America. *Influenza and Other Respiratory Viruses*, 3(5):215–222.

Price, D. J., Shearer, F. M., Meehan, M. T., McBryde, E., Moss, R., Golding, N., Conway, E. J., Dawson, P., Cromer, D., Wood, J., et al. (2020). Early analysis of the Australian COVID-19 epidemic. *ELife*, 9:e58785.

Riley, S., Fraser, C., Donnelly, C. A., Ghani, A. C., Abu-Raddad, L. J., Hedley, A. J., Leung, G. M., Ho, L.-M., Lam, T.-H., Thach, T. Q., et al. (2003). Transmission dynamics of the etiological agent of SARS in Hong Kong: impact of public health interventions. *Science*, 300(5627):1961–1966.

Tuite, A. R. and Fisman, D. N. (2020). Reporting, epidemic growth, and reproduction numbers for the 2019 novel coronavirus (2019-nCoV) epidemic. *Annals of Internal Medicine*, 172(8):567–568.

Tuite, A. R., Greer, A. L., Whelan, M., Winter, A.-L., Lee, B., Yan, P., Wu, J., Moghadas, S., Buckeridge, D., Pourbohloul, B., et al. (2010). Estimated epidemiologic parameters and morbidity associated with pandemic H1N1 influenza. *CMAJ*, 182(2):131–136.

van den Driessche, P. and Watmough, J. (2002). Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission. *Mathematical Biosciences*, 180(1-2):29–48.

van Gageldonk-Lafeber, A. B., van der Sande, M. A., Meijer, A., Friesema, I. H., Donker, G. A., Reimerink, J., van Beest, M. R.-D. R., Prins, J. M., Isken, L., Schellevis, F. G., et al. (2012). Utility of the first few 100 approach during the 2009 influenza A (H1N1) pandemic in the Netherlands. *Antimicrobial Resistance and Infection Control*, 1(1):1–7.

Vegvari, C., Abbot, S., Ball, F., and et al. (2021). Commentary on the use of the reproduction number R during the COVID-19 pandemic. *Statistical Methods in Medical Research*.

Vink, M. A., Bootsma, M. C. J., and Wallinga, J. (2014). Serial intervals of respiratory infectious diseases: a systematic review and analysis. *American Journal of Epidemiology*, 180(9):865–875.

Wang, W. and Ruan, S. (2004). Simulating the SARS outbreak in Beijing with limited data. *Journal of Theoretical Biology*, 227(3):369–379.

White, L. F., Moser, C. B., Thompson, R. M., and Pagano, M. (2021). Statistical estimation of the reproductive number from case notification data. *American Journal of Epidemiology*.

White, L. F. and Pagano, M. (2008). A likelihood-based method for real-time estimation of the serial interval and reproductive number of an epidemic. *Statistics in Medicine*, 27:2999–3016.

World Health Organization and others (2020). The First Few X cases and contacts (FFX) investigation protocol for coronavirus disease 2019 (COVID-19). Technical report, World Health Organization.

Zhao, S., Lin, Q., Ran, J., Musa, S. S., Yang, G., Wang, W., Lou, Y., Gao, D., Yang, L., He, D., et al. (2020). Preliminary estimation of the basic reproduction number of novel coronavirus (2019-nCoV) in China, from 2019 to 2020: A data-driven analysis in the early phase of the outbreak. *International Journal of Infectious Diseases*, 92:214–217.