EXPLORING THE NUANCES OF $R_0$: EIGHT ESTIMATES AND APPLICATION TO 2009 PANDEMIC INFLUENZA

A PREPRINT

Shannon Gallagher  
Biostatistics Research Branch, Division of Clinical Research  
National Institute of Allergy and Infectious Diseases  
Rockville, MD 20852  
shannon.gallagher@nih.gov

Andersen Chang  
Department of Statistics Rice University  
Houston, TX 77005  
atc7@rice.edu

William F. Eddy  
Department of Statistics & Data Science  
Carnegie Mellon University  
Pittsburgh, PA 15213

March 25, 2020

ABSTRACT

For nearly a century, the initial reproduction number ($R_0$) has been used as a one number summary to compare outbreaks of infectious disease, yet there is no ‘standard’ estimator for $R_0$. Difficulties in estimating $R_0$ arise both from how a disease transmits through a population as well as from differences in statistical estimation method. We describe eight methods used to estimate $R_0$ and provide a thorough simulation study of how these estimates change in the presence of different disease parameters. As motivation, we analyze the 2009 outbreak of the H1N1 pandemic influenza in the USA and compare the results from our eight methods to a previous study. We discuss the most important aspects from our results which effect the estimation of $R_0$, which include the population size, time period used, and the initial percent of infectious individuals. Additionally, we discuss how pre-processing incidence counts may effect estimates of $R_0$. Finally, we provide guidelines for estimating point estimates and confidence intervals to create reliable, comparable estimates of $R_0$.

Keywords reproduction number · 2009 pandemic influenza · SIR · compartment models

1 Introduction

What has been called “arguably the most important quantity in the study of epidemics,” $R_0$ (by convention pronounced “R-naught”), the initial reproduction number remains an important quantity to estimate to assess the severity of infectious diseases. As defined by [Anderson and May](1992), $R_0$ is the “the average number of secondary infections produced when one infected individual is introduced into a host population where everyone is susceptible.” Often understood to be synonymous with the severity of infection, $R_0$ is still used to inform both professionals and lay people alike. For example, the Wall Street Journal in February 2020 reported an estimate of $R_0$ before the impending outbreak of COVID-19, a novel coronavirus, in the West (Craven McGinty, 2020). Because of this reliance upon this estimate, it is more important than ever to produce accurate and reliable estimates of this quantity.

In some ways, $R_0$ summarizes the entire outbreak of a disease. $R_0$ is used to assess whether a disease outbreak will occur and its severity. Additionally, $R_0$ describes what percentage of the population needs to be vaccinated to avoid such an epidemic, roughly $1 - R_0^{-1}$; is used to estimate the final size of the total number of infected individuals; and is related to the probability of observing an outbreak under the same conditions (Anderson & May 1992, Britton 2010). Despite a clear definition of $R_0$, epidemiologists have struggled to create a standard estimator for $R_0$ (Hethcote 2000).
One major issue in estimating $R_0$ is that the quantity is a property of the model, meaning that $R_0$ is dependent not only on the usual noise that comes with statistical modeling but also on a variety of assumptions on how researchers assume a disease is transmitted through a population (Brown, Oleson, & Porter, 2016; Diekmann, Heesterbeek, & Roberts, 2009). The consequences of this is that despite being nominally the same, $R_0$ may have a different interpretation depending on the model used. As an example, there are two common types of epidemic frameworks to describe the transmission of infectious disease; SIR and SEIR models. Here, ‘S’ stands for susceptible, ‘I’ for infectious, ‘R’ for recovered or removed, and ‘E’ for exposed (i.e. infected but not yet infectious). In both models, $R_0$ is defined in the same manner, namely the infection rate over the recovery rate, $R_0 = \frac{\gamma}{\beta}$. However, the two estimates are not directly comparable to one another. An analogy is comparing coefficient $\beta_1$ that appears in two linear regression models, where the second model is nested in the first, as shown in Eq. (1) and (2).

$$E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

(1)

$$E[Y] = \beta_0 + \beta_1 X_1.$$  

(2)

In both models, estimates of $\beta_1$ can indicate whether $X_1$ is a significant variable or not. However, in Eq. (1), $\beta_1$ can only be interpreted in the context of $X_2$ whereas the interpretation $\beta_1$ in Eq. (2) has no explicit connection to $X_2$.

A number of papers have been written about the difficulties and nuances involved in estimating $R_0$ (see Diekmann et al., 2009; Hethcote, 2000; Lt. Blakeley et al., 2011; Van den Driessche, 2017). Our study differs from those previous through a study of different methods to estimate $R_0$, emphasis on the data format, and sensitivity to certain disease parameters. However, even in the same model framework (e.g. SIR models), estimates may differ from one another. Difference in estimates can result from mathematical versus statistical methods (i.e. solving for $R_0$ versus estimating $R_0$), the number of observations used to estimate $R_0$ (e.g. time steps), population size, initial SI ratio, and model specification. These are all issues for producing point estimates, to say nothing of issues related to estimating confidence intervals (CI). Moreover, there is the important issue of the data itself. Methods to estimate SIR curves typically assume the observed number of Susceptible, Infectious, and Recovered ($S(t)$, $I(t)$, $R(t)$) are available at a number of time steps, but typically those values are unobserved, and we, instead, observe only newly reported cases in a given time interval.

In this paper, we focus on highlighting the three following aspects of $R_0$: 1) variety in estimators, 2) sensitivity to certain disease parameters, and 3) how the difference between ideal and actual data can influence our estimates of $R_0$. We limit our estimates to eight estimates that are constructed to work within the SIR framework. Moreover, we consider Bayesian models to be outside the scope of our paper. Our results are shown in both a simulation study as well as an application to 2009 pandemic influenza.

The rest of this paper is organized as follows. In Section 2, we introduce the deterministic SIR model and stochastic variations of it. We then describe eight different estimates of $R_0$ in Sections 2.1 to 2.8. Following that in Section 3, we describe the data and nuances thereof of both our simulation data and the 2009 pandemic influenza data. In Section 4, we describe the results of the simulation study the 2009 pandemic influenza application. Finally, in Section 5, we discuss our recommendations and conclusions from our analysis.

2 Methods

The deterministic model. In order to discuss our eight methods to estimate $R_0$ in the SIR model, we first need to define the SIR model. The SIR model introduced by Kermack and McKendrick (1927) is a compartment model, where individuals move from susceptible, to infectious, and finally recovered states (compartments). We study the SIR model without vital dynamics (e.g. no birth and death into or out of a population). We make five essential assumptions: 1) the compartments are discrete and have no overlap, 2) transition of individuals is described by a set of known equations, possibly dependent on an unknown parameter, 3) the populations mix homogeneously, 4) the number of individuals in each compartment at time $t = 0$ is known, and 5) the law of mass action is obeyed. The last of these conditions is a property borrowed from chemistry which says that the mass of the product per unit time is proportional to the mass of the reactants (Lotka, 1920). In epidemiology, this means that the proportion of new infections per unit time is proportional to the current number of susceptible (Anderson & May, 1992).

In this SIR model, (displayed graphically in Figure 1), the total number of individuals ($N$) is constant. Adaptations of the SIR model can include birth and death rates, which may correspondingly change the derivation of $R_0$ (for further discussion. Recall, $\beta$ is the average infection rate and $\gamma$ is the average recovery rate. We assume $\beta$ and $\gamma$ are positive. The movement of individuals from one compartment to another is represented through the ordinary differential equations below. For the remainder of this paper, we use $X$, $Y$, and $Z$ to denote the number of individuals in the $S$, $I$, $R$
and $R$ compartments, respectively, to avoid any confusion with $R_0$.  

$$
\begin{align*}
\frac{dX}{dt} &= -\frac{\beta XY}{N} \\
\frac{dY}{dt} &= \frac{\beta XY}{N} - \gamma Y, \\
\frac{dZ}{dt} &= \gamma Y.
\end{align*}
$$

(3)

In words, susceptible individuals become infected at a rate that is proportional to the percentage of infected individuals multiplied by $\beta$, the infection rate, and the number of susceptible individuals. Infectious individuals recover at a rate of $\gamma$ multiplied by the number of infected individuals. Whenever we refer to the “original SIR model” in this paper, we mean the set of equations given in Eq. (3) along with known initial state counts, $(X(0), Y(0), Z(0))$.  

Figure 1: Depiction of a SIR model where $X = S, Y = I,$ and $Z = R$. One can only be infected once by the disease in this model. Individuals begin in the $X$ state as susceptible individuals, possibly become infectious and move to the $Y$ state, and finally possibly recover in the $Z$ state. There are no births or deaths in this SIR model.

An outbreak occurs if the rate of change of infectious individuals is positive, $\frac{dY}{dt} > 0$, or equivalently, $\frac{\beta}{\gamma} > \frac{N}{X}$, since the sum of the derivatives is zero as the population $N$ is constant. That is, an outbreak occurs if the rate of new infections is greater than the rate of recovery. So as long as the number of initially susceptible individuals is large compared to the total population, i.e. $\frac{X}{N} \approx 1$, then an outbreak will occur if $R_0 > 1$, where

$$
R_0 \overset{\text{def}}{=} \frac{\beta}{\gamma}.
$$

The stochastic model. To incorporate randomness into the model, we add noise to the compartments, namely,

$$
\begin{align*}
\hat{X}(t) &= X(t) + \epsilon_{X,t} \\
\hat{Y}(t) &= N - \hat{X}(t) - \hat{Z}(t) \\
\hat{Z}(t) &= Z(t) + \epsilon_{Z,t}
\end{align*}
$$

(4)

The “hats” in Equation (4) are used to distinguish random variables from the “true,” deterministic model without hats (e.g. $\bar{X}$ vs. $X$). We are assuming the observations are generated from the deterministic ODEs (possibly converted to discrete time) presented in Equation (3) along with time and compartment dependent noise $\epsilon_{X,t}$ and $\epsilon_{Z,t}$. Since $N$, the total population, is constant, then $Y$ is adjusted accordingly. An advantage to adding noise to the $X$ and $Z$ compartments as opposed to the $X$ and $Y$ or the $Z$ and $Y$ compartments is that we can more readily enforce the $X$ and $Z$ compartments to be monotonic in $t$. This is useful as we expect the number of susceptible to be non-decreasing and the number of recovered to be non-increasing in an SIR disease-transmission process.

In summary, when we discuss estimators for $R_0$ for the SIR model, we mean to say we are forming an estimator of $R_0$ from the given set of data in Eq. (4)

$$
\text{Data} = \left\{ \left( \hat{X}(t) = x(t), \hat{Y}(t) = y(t), \hat{Z}(t) = z(t) \right) : t = t_0, t_1, \ldots, t_T \right\},
$$

(5)

where $m$ is a function of the data. We note that our SIR formulation does not include other population processes, such as vital dynamics (e.g. birth and death rates). This will be true throughout our analysis, and we will discuss the implications of this in a later section.

**Review of eight methods.** In the following subsections, we review eight methods to estimate $R_0$, that are all of the form shown in Eq. (4), that is, an estimate of $R_0$ that is a function of XYZ data. We examine both unspecified distribution and specified distribution-based estimates. By unspecified-distribution, we refer to estimates that do not explicitly identify the distributions of random processes. For example, we describe a method that estimates $R_0$ through minimizing the $L_2$ norm of the data and expected model estimates. In contrast, distribution-based estimates refer to the methods that do explicitly identify a generative model to incorporate noise. For example, we describe an estimate based on the chain Binomials described by Abbey (1952), which describes the transmission process via Binomial draws.
In the below sections, we describe the eight methods, but before we describe the methods, we would like to comment on some of the methods we omitted from this analysis. Originally, we included 19 unique methods to estimate \( R_0 \). Many of these 19 methods were similar enough to one another to be combined into one group. For example, some methods used a Binomial variable to model the transmission of new infections whereas others used a Poisson variable but were otherwise the same. A number of these similar methods were hierarchical models where the initial transmission and recovery were modeled with parameters and hyperparameters were then placed upon those parameters. While we recognize the importance and utility of such models, we decided to focus on non-hierarchical models and consider these other models to be variants on these simplified, core models.

Other omissions were made because upon close inspection, the method presented did not fit the SIR paradigm. For example, branching processes such as the Galton-Watson process (as seen in Farrington, Kanaan, and Gay (2003)) are of historical importance in epidemiology. However, such processes presume a limitless population size, and the recovery process is not explicitly modeled. Despite this, many parts of the branching process can be seen in our Markov Chain model.

Another notable omission is that of the individual-based (IBM) or agent-based model (ABM). In these models, individuals or agents interact with each and their environment over time to produce difficult-to-calculate effects (Blower & Chou, 2004; Epstein, 2007). However, IBMs and ABMs rarely satisfy our assumption of homogeneity among individuals in the same compartment, which we require in the eight methods. We hope to study \( R_0 \) estimates in such models in future work, because IBMs and ABMs present a fairly unique way to estimate \( R_0 \), as we can directly count who infected whom at each time point.

2.1 Exponential Growth (EG)

Wallinga and Lipsitch (2007) report that the effective reproduction number \( R_0 \) and hence the initial reproduction number \( R_0 \) may derived using the fact that infection “counts increase exponentially in the initial phase of an epidemic.” We then have to estimate \( r \), the per capita change in the number of new cases per unit of time and \( \omega \) the serial interval, the distribution of time between a primary and secondary infection. Then, we have

\[
R_0 = \exp(r\omega)
\]  

Equation (6) is derived from a demographic view using the Lotka-Euler survival equations which come from the fields of demography, ecology, and evolutionary biology. Sometimes, in epidemiology, the first order Taylor series expansion of Eq. (6) is instead used. This method is often taught as a way to introduce the concept of \( R_0 \) and a variant can be seen in Nishiura, Chowell, Safan, and Castillo-Chavez (2010). Commonly, we assume that the serial interval \( \omega \) is the same as infection duration, \( \omega \approx \gamma - 1 \) and during the ‘initial growth phase’ that the number of susceptibles is equal to

\[
X(t) = X(0)e^{-rt}.
\]

To estimate \( R_0 \), we then must estimate \( r \),

\[
\hat{r} = \arg \min_r \sum_{t=t_{\min}}^{t_{\max}} (X(t) - x(t))^2
\]

\[
\hat{R}_0 = 1 + \frac{\hat{r}}{\gamma}.
\]

The advantages to this method are that it relies only on estimates of the number of susceptibles. However, it assumes exponential growth during the initial growth phase and so relies on knowing when that initial growth phase occurs. Nishiura et al. (2010) provide guidelines of when such a method should be used because of the initial growth assumption. This method also relies on having some estimate of \( \gamma \) available. Tuning parameters for this method include \( t_{\max} \), the maximum time point of exponential growth in the number of infections is observed (or the corresponding exponential decay in the number of susceptibles), possibly \( t_{\min} \), the minimum time point, and \( \gamma \) (if it is not treated as a random variable). We choose that \( t_{\min} = 0 \) and \( t_{\max} \) is the time corresponding to the maximum number of infectious of the observed data.

There are a number of adjustments one can make to this method. For example, Wallinga and Lipsitch (2007) detail a way to estimate \( R_0 \) through treating \( \omega \) as a random variable. Some variations of this method assume that \( r \) has its own distribution (Obadia, Haneef, & Boëlle, 2012).

2.2 Ratio Estimator (RE)

Our second approach to estimate \( R_0 \) in the SIR model is to minimize the joint mean square error for the data collected at each time point and use the plug-in estimator found in Equation 8. In particular, we find:
\[(\hat{\beta}, \hat{\gamma}) = \arg\min_{\beta, \gamma} \sum_{t} \left[ (x(t) - X(t; \beta, \gamma))^2 + (y(t) - Y(t; \beta, \gamma))^2 + (z(t) - Z(t; \beta, \gamma))^2 \right] \]

Then the ratio estimator (RE) estimate for \( R_0 \) is given by Equation 8:

\[ \hat{R}_0 = \frac{\hat{\beta}}{\hat{\gamma}} \]  

The estimates resulting from this method are often found via a grid search of \( \beta \) and \( \gamma \) values to minimize the sum square errors. The estimates can also be found using more sophisticated optimization algorithms. Since we typically cannot explicitly write down the partial derivatives of \( S, I, \) and \( R \) with respect to \( \beta \) and \( \gamma \), we can only use minimization algorithms which do not rely on an explicit gradient, such as Nelder-Mead simplex minimization. RE and similar variants have been used to estimate \( R_0 \) in Majumder et al. (2016).

A reasonable question one may ask is why we use the \( L_2 \)-norm as opposed to \( L_1 \) or some other similarity score. One answer is that if we were to assume independent and identically distributed Gaussian noise, then the \( L_2 \)-norm would be equivalent to the maximum likelihood estimation. Another answer is that \( L_2 \) is a continuously differentiable function and hence is easier to compute, which allows for sensitivity analysis to be conducted more easily. Finally, the RE estimate can be found without writing down explicit assumptions about the noise within the model. However, we cannot guarantee properties of consistency or convergence to a known distribution without more explicit assumptions.

For the most part, there are not many adjustments that can be made to this method because we do not explicitly specify a stochastic model for how the data are generated. As such, we have no nuisance parameters or tuning parameters that we need to adjust when fitting our model.

2.3 Reparameterized Ratio Estimator (rRE)

In the previous method (Section 2.2), we estimated \( \beta \) and \( \gamma \) and then estimated \( R_0 \). However, it is possible to directly estimate \( R_0 \) if we reparameterize the ODEs in Equation (3) directly with \( R_0 \) and \( \gamma \), using the relation \( R_0 = \frac{\hat{\beta}}{\hat{\gamma}} \). We find

\[ \left( \hat{R}_0, \hat{\gamma} \right) = \arg\min_{R_0, \gamma} \sum_{t} \left[ (x(t) - X(t; R_0, \gamma))^2 + (y(t) - Y(t; R_0, \gamma))^2 + (z(t) - Z(t; R_0, \gamma))^2 \right] \]

We use the \( \hat{R}_0 \) directly from the above estimation problem, which again can be found with a grid search or another optimization process. It may be surprising that this method can lead to different results than simply using the ratio estimator. In our simulation study, we show that we have differences in both the mean and variance of the two estimates of \( R_0 \). The difference in mean can be attributed to the practical issue of finding the arguments that minimize the objective function (e.g. \( R_0 \) is typically not near the boundary of zero whereas \( \beta \) and \( \gamma \) are more likely to be closer to that boundary). The difference in variance is due to both the difference in mean (as the mean estimate is used in the variance estimate) and also because RE estimates the variance by first estimating the Hessian of the ‘true’ function numerically and then applying the delta method to estimate the variance of \( R_0 \). On the other hand, in rRE only the Hessian needs to be estimated.

Like RE, while we have no explicit assumptions on the noise. Likewise, the same difficulties in sensitivity analysis arise as for the RE. We are unaware of any attempts in the literature to directly estimate \( \beta \) and \( \gamma \) in RE and then taking the ratio to estimate \( \hat{R}_0 \).

2.4 Log Linear (LL)

Harko, Lobo, and Mak (2014) reduce the SIR model with two ODEs and one constraint to one ODE and one constraint. From this, we can derive the following,

\[ \log \left( \frac{X(t)}{X(0)} \right) = -R_0 \frac{Z(t)}{N}. \]  

We can regress the left hand side in Eq. 9 on \( Z(t) \), the number of recovered individuals, using least squares to estimate the coefficients with \( R_0 \) as the coefficient of \( Z(t)/N \) to obtain an estimate of \( \hat{R}_0 \),

\[ \hat{R}_0 = -\frac{\sum_{t=0}^{T} \log \left( \frac{X(t)}{X(0)} \right)}{\sum_{t=0}^{T} \frac{Z(t)}{N}}. \]
This method says that for a one percent increase in the number of recovered individuals over time period $\Delta$, we expect the ratio of the number of previous susceptibles to number of new susceptibles to increase by $e^{.01R_0}$,

$$\frac{X(t)}{X(t + \Delta)} = e^{.01R_0}.$$  

An advantage of this method is an alternative interpretation of $R_0$: in terms of the ratio of the old number of susceptibles to new number of susceptibles for a specific amount of time. Another advantage is that this method is easily implemented using any linear modeling software. To our knowledge, the log-linear model has not yet been applied to a data setting.

The method also has disadvantages in that it is assumes the expected path of the individuals exactly follows the Kermack and McKendrick differential equations, which makes this model choice very sensitive to the disease transmission process. As a result, we currently do not know of any variants for this method.

### 2.5 Markov Chain (MC)

A natural approach to epidemic modelling is that of Markov chains (MC), since it is assumed an individual’s next state is only dependent on its current state and the current states of other individuals. Much work has been done over the years in this specific field including asymptotic behavior, continuous time MC, confidence intervals, and more (Daley, Gani, & Gani, 2001; Gani & Yakowitz, 1995; Jacquez & O’Neill, 1991). We present one simple instance of the model, the discrete time case, which traces its origin back to the Reed-Frost model (Abbey, 1952).

In this model, the number of susceptibles at the next step, $\tilde{X}(t + 1)$, has a Binomial distribution based on the contacts with the current number of infectious, $\tilde{Y}(t)$ and the current number of susceptibles. This is based on estimating $\alpha$, the probability of coming into contact and subsequently becoming infected by a single infectious individual.

Unlike some of the other methods, the Reed-Frost Chain Binomial has a very specific form of $Y(t)$, the number of infectious at time $t$. We assume that that the incidence at time $t$ is equal to the number of infectious at at time $t$,

$$\tilde{X}(t) = \tilde{X}(t-1) - \tilde{Y}(t)$$

$$\tilde{Y}(t)|\tilde{X}(t-1), \tilde{Y}(t-1) \sim \text{Binomial} \left( \tilde{X}(t-1), 1 - (1 - \alpha)^{\tilde{Y}(t-1)} \right)$$

$$\tilde{Z}(t) = \tilde{Z}(t-1) + \tilde{Y}(t)$$

The full likelihood for $\alpha$ in this model, namely,

$$\mathcal{L}(\alpha; \text{Data}) \propto \prod_{t=0}^{T-1} \left( 1 - (1 - \alpha)^{\tilde{Y}(t)} \right)^{\tilde{X}(t+1)} \left( 1 - (1 - \alpha)^{\tilde{Y}(t)} \right)^{\tilde{X}(t)-\tilde{X}(t+1)}.$$  

Maximizing the likelihood yields an estimate for $\alpha$,

$$\hat{\alpha} = \arg\max_{\alpha} \mathcal{L}(\alpha; \text{data}).$$

An approximate estimate of the reproduction number is then,

$$\tilde{R}_0 = \log \left( \frac{1}{1 - \hat{\alpha}} \right).$$

This method typically allows for more than just the reproduction number to be estimated. Through recursion, one can estimate the probability of having a given number of susceptibles and infected at each time step, and hence the entire probability distribution may be known. The advantages of this model are its simplicity of interpretation and ability to generate an estimate for the probability distribution for $(X(t), Y(t))$.

Essentially, in this model we assume that all infectious cases have recovered by the next time step or equivalently, that $\gamma = 1$. As a result, it is likely that this method will overestimate $R_0$ except in cases where, for example, infectiousness does not appear until symptom onset and at discovery of symptoms the patient is immediately treated and no longer infectious.

### 2.6 Likelihood-Based Estimation (LBE)

A common way used to estimate epidemiological quantities is through likelihood based estimation. We saw an example of this in the previous section with MC estimation. However, as we will see in the results, this particular model does not
seem to fit the data well. As a result, many researchers specify different models for the generation of SIR data; estimate the likelihood through direct calculation or approximate the likelihood through simulation such as iterated filtering used in “plug-and-play” and finally derive point estimates and CIs from this likelihood estimate $R_0$ (Abbey, 1952; Bhadra et al., 2011; Forsberg White & Pagano, 2008; He, Ionides, & King, 2009). Here, we examine one model that has been used to model the discrete-time SIR: a binomial chain with probabilities based on the deterministic SIR model.

The discrete time XYZ model is as follows, where we have replaced $d$ with $\Delta$,

$$\begin{aligned}
\frac{\Delta X}{\Delta t} &= -\beta YX \\
\frac{\Delta Y}{\Delta t} &= \frac{\beta YX}{N} - \gamma Y \\
\frac{\Delta Z}{\Delta t} &= \gamma Y
\end{aligned}$$

Denote $\hat{X}(t), \hat{Y}(t)$, and $\hat{Z}(t)$ to be the observed number in each compartment and $X(t), Y(t)$, and $Z(t)$ to be the true, underlying model at time $t$. The stochastic model is given by

$$\begin{aligned}
X(t+1)|\hat{X}(t), \hat{Z}(t) &= \hat{X}(t) - \epsilon_{X,t} \\
Y(t+1)|\hat{X}(t), \hat{Z}(t) &= N - \hat{X}(t+1) - \hat{Z}(t+1), \\
Z(t+1)|\hat{X}(t), \hat{Z}(t) &= \hat{Z}(t) + \epsilon_{Z,t}
\end{aligned}$$

with $\hat{X}(0), \hat{Y}(0)$, and $\hat{Z}(0)$ known. The terms $\epsilon_{X,t}$ and $\epsilon_{Z,t}$ in Equation (4) incorporate random behavior into the model and we constrain $\beta, \gamma \in [0,1]$. Here we choose them to have the following distributions,

$$\begin{aligned}
\epsilon_{X,t}|\hat{X}(t), \hat{Z}(t) &\sim \text{Binomial} \left( \hat{X}(t), \frac{\beta \hat{Y}(t)}{N} \right) \\
\epsilon_{Z,t}|\hat{X}(t), \hat{Z}(t) &\sim \text{Binomial} \left( \hat{Y}(t), \gamma \right)
\end{aligned}$$

If we assume that the next time step is dependent only on the current time step, then the likelihood is

$$\mathcal{L}(\beta, \gamma; \text{Data}) = \prod_{t=0}^{T-1} \left[ \frac{x_t}{\epsilon_{X,t}} \right]^{\epsilon_{X,t}} \left( \frac{\beta y(t)}{N} \right)^{\epsilon_{X,t}} \left( 1 - \frac{\beta y(t)}{N} \right)^{x(t) - \epsilon_{X,t}} \left[ \frac{y(t)}{\epsilon_{Z,t}} \right]^{\epsilon_{Z,t}} \left( 1 - \gamma \right)^{y(t) - \epsilon_{Z,t}}.$$ 

The estimate for $R_0$ is then

$$\hat{R}_0 = \frac{\hat{\beta}}{\hat{\gamma}}$$

The two main differences between this model and MC are that 1) LBE looks at the joint likelihood of the XYZ model, whereas MC assumes the previous infectious individuals all recover from one time step to the next and 2) the difference in the structure of the transition probability. A major advantage of this model easily modified and can be used to as a basis for more sophisticated models, and can be easily estimated in available software such as pomp (King, Nguyen, & Ionides, 2016). Disadvantages include maximizing the likelihood may be more time consuming than other methods due to the need to approximate $X(t)$ and $Z(t)$ for many different values of $\beta$ and $\gamma$ and that this model is sensitive to the underlying assumptions of how the disease is stochastically transmitted. There are no tuning or nuisance parameters that need to be estimated.

2.7 Incidence to Prevalence (IPR)

The incidence to prevalence ratio (IPR), described by [Nishiura and Chowell, 2009], is another intuitive method to estimate $R_0$. It incorporates some of the most basic epidemiological quantities: incidence and prevalence. In terms of data from the SIR model, the incidence is approximately the number of new infectious for a given time step, $J(t) \approx (X(t + 1) - X(t))$, and the IPR is the ratio of incidence to prevalence, $\text{IPR}(t) = \frac{J(t)}{\hat{Y}(t)}$. This method assumes that we have some prior knowledge about $\gamma$, the recovery rate. Thus we use as our estimate,

$$\hat{R}_0 = \text{IPR}(t^*) \cdot \frac{1}{\hat{\gamma}}$$
where $t^*$ is the time at which the maximum prevalence is reached.

Here we assume that the time step is small enough to approximate the incidence. The advantage of this method is that incidence data is generally readily available as is prevalence data for certain diseases. However, as one is required to have prior knowledge about $\gamma$, it may be easier to directly estimate $R_0$ with one of the many other methods described that does not require prior knowledge. Again, we are using only one time point to estimate $R_0$. This model places no assumptions on the noise. Here, $\gamma$ may be treated as a tuning parameter.

2.8 Linear model approximation (LMA)

In the past decade, non-parametric and semi non-parametric methods have been increasingly used to (quite successfully) predict the prevalence of a disease such as the methods by Brooks, Farrow, Hyun, Tibshirani, and Rosenfeld (2015) used in Center of Disease Control and Prevention’s yearly influenza prediction contest. This success in prediction prompted us to ask whether success in non-parametric curve fitting could transfer to estimating $R_0$.

The idea is as follows: assume $(\hat{X}(t), \hat{Y}(t), \hat{Z}(t)) = (f(t), g(t), h(t)) + \epsilon_k$ where $f, g, h$ are functions such that $f'(t), g'(t), h'(t)$ exist for all $t \in (0, T)$. Then a rough estimate of $R_0$ is

$$\hat{R}_0 \approx -\frac{f'(0)}{h'(0)} \times \frac{N}{\hat{X}(0)}$$

where $\hat{f}$ and $\hat{h}$ are smooth estimates of the actual underlying distributions. For example, we could have $\hat{f}$ be a polynomial function in $t$.

The linear approximation of the Kermack and McKendrick SIR model has been studied previously [Chen & Li, 2009; Hu, Teng, & Long, 2014]. We extend these linear models to polynomial models. Alternatively, one can think of this method as a naive way to approximate the $S$, $I$, and $R$ curves. We use this method to estimate $R_0$. Specifically, we estimate two polynomials in $t$ with degree $K$ using least squares to find the coefficients $\{(\hat{x}_k, \hat{z}_k)\}_{k=1,...,K}$ and use the resulting polynomial coefficients to estimate the derivatives, $\hat{X}'(t)$ and $\hat{Y}'(t)$. An estimator for $R_0$ is derived from the ODEs in Equation (3).

$$\hat{R}_0 = -\frac{\hat{X}'(0)}{\hat{Z}'(0)} \times \frac{N}{\hat{X}(0)}.$$  \hspace{1cm} (12)

Here, $K$ is arbitrary and should be selected using some criterion such as AIC. Besides optionally deciding on the degree of polynomials to fit, this model is simple to implement. The time $t = 0$ is used to best capture the initial outbreak. An advantage of using this estimation method is that it is simple to implement using any linear modelling software.

A prominent disadvantage of this method (and in general using semi or non-parametric curve estimates to estimate $R_0$) is that we implicitly assume that the underlying disease curves are mechanistic in nature, specifically of the mechanism of disease transmission indicated by the SIR model. As such, it may make no sense in the first place to estimate the curves in a non-parametric manner, when ultimately we are using those estimates to estimate a disease parameter. Moreover, if the disease transmission process is not truly described by the SIR model, then our derivation of $R_0$ may be meaningless and so we may be estimating the wrong quantity of interest.

Despite this prominent disadvantage, we include this model in our testing because non-parametric methods have been successfully used recently for the prediction of disease models, and it is a natural question of whether we can ‘plug and chug’ the curve estimates to obtain a reliable estimate of $R_0$. If the estimates are reliable (or reliable enough), then we can use curves for both prediction and inference.

Methods Summary. Sections 2.1-2.8 describe how to form an estimate of $R_0$ for eight different methods. The eight methods are summarized in Table 1. All of our methods explicitly describe point estimates of $R_0$. In addition, we report 95% CI to go along with every point estimate. These 95% CI are determined by appropriate use of the Delta Method as described in Wasserman (2004).

3 Data

An important and sometimes glossed-over issue in estimating parameters in SIR models (and compartment models in general) is how the data is "pre-processed" into a format that the mentioned methods can use. The methods for estimating $R_0$ in Section 2 rely on observations of $(x(t), y(t), z(t))$, the number of individuals in the susceptible, infectious, and recovered compartments at different time points. However, such observations only occur in fairly contrived situations such as small, closed populations where infection time and recovery time are almost immediately
Table 1: Table of methods used. The heading “Specified distribution?” refers to whether a statistical distribution is explicitly specified in the model. The methods are described in Sections 2.1-2.8.

| Abbr. | Method                  | Specified distribution? | Reason for inclusion         |
|-------|-------------------------|--------------------------|------------------------------|
| EG    | Exponential Growth      | No                       | Naive approximation          |
| RE    | Ratio Estimator         | No                       | History, simplicity, optimization |
| rRE   | Reparamatrized RE       | No                       | Slight adjustment to known model |
| LL    | Log Linear              | No                       | Slight adjustment to known model |
| MC    | Markov Chain            | Yes                      | History, plausibility        |
| LBE   | Likelihood Based Estimation | Yes                    | Common method of estimation  |
| IPR   | Incidence to Prevalence | No                       | Typical data                |
| LMA   | Linear Model Assumption | No                       | Non-parametric approach      |

known. What is far more common is, instead, to have a fairly good estimate of the population size $N$ and the new case counts (incidence) $j(t)$ on a daily, weekly, or monthly basis. From there, the data may be pre-processed into a usable form that the above methods may use. For example, if $N$ is large, the number of susceptibles can be reasonably estimated from the case counts as $x(t) = N - \sum_{s=0}^{t} j(s)$, the population size minus the cumulative sum of the incidence over the time period. However, it is typically more difficult to disentangle $y(t)$ and $z(t)$.

One approach to separate the number of infectious from the number of recovered is to employ the use of a known estimate of $\gamma$, the rate of recovery or equivalently $\gamma^{-1}$, the average time to recovery for an infectious individual, as seen in Towers and Feng (2009). We assume that the number of infectious is the previous number of infectious, plus the new cases and minus the new recovered cases. In fact, the above is a discrete approximation of the SIR model where $j(t)$ is the new cases flowing in, $y(t-1)$ is the previous number of cases and $\gamma y(t-1)$ is the number of cases flowing out of the infectious state and so are subtracted. The transformation from incidence to the XYZ values are given by the following equations,

$$
\begin{align*}
  x(t) &= N - \sum_{s=0}^{t} j(s) \\
  y(t) &= j(t) + y(t-1)(1-\gamma) \\
  z(t) &= N - x(t) - y(t),
\end{align*}
$$

(13)

The consequence of pre-processing the data is that we are imposing SIR (XYZ) structure on our observed data. This will almost certainly have an effect of estimating $\gamma$ and thus $R_0$, but this pre-processing step remains a common approach (as seen recently in Wang et al. (2020)). Other approaches to pre-processing include treating the XYZ compartment values as latent variables or through data augmentation (Fintzi, Cui, Wakefield, & Minin, 2017; He et al., 2009). We find these other approaches to be useful and valuable but do not analyze them here.

In our simulation study, we generate and fit our methods on data of the form $(x(t), y(t), z(t))$, i.e. the ideal situation where we know the counts of all three compartments. In contrast, the 2009 pandemic influenza data uses an approximation of the incidence (weighted influenza like illness). In that case, we use the pre-processing steps described above in Eq. (13) to format the data into a usable SIR format. Because the data is approximated, we perform sensitivity analysis on $\gamma$.

3.1 Simulation data

In this section, we give the details of how we generate our simulation data. Since we want to compare the eight methods to one another, we strive to compare data with both process and measurement error, where the former refers to the error from the inherent variability in transmission model, and the latter to observation error. In all of these simulations, we first generate data from the models under known functions $(f, g)$. We then add noise drawn from known distributions $(\epsilon_{X,t}, \epsilon_{Y,t})$ to each time point,

$$
\begin{align*}
  \hat{X}(t) &= f(t) + \epsilon_{X,t} \\
  \hat{Y}(t) &= N - \hat{X}(t) - \hat{Z}(t) \\
  \hat{Z}(t) &= N + \epsilon_{Z,t}.
\end{align*}
$$

(14)
To maintain a constant population, we adjust the recovered compartment \( \hat{Y}(t) \) after \( \hat{X}(t) \) and \( \hat{Z}(t) \) are determined. We simulate data that has a combination of process error and measurement error. We assume that the initial values \((X(0), \hat{Y}(0), \hat{Z}(0))\) are known.

**Simulation data generation**

\[
\begin{align*}
    & f(t) = \hat{X}(t - 1) \\
    & g(t) = \hat{Z}(t - 1) \\
    & \epsilon_{X,t} | \hat{X}(t - 1), \hat{Z}(t - 1) = \alpha_{X,t} + \tau_{X,t} \\
    & \epsilon_{Z,t} | \hat{X}(t - 1), \hat{Z}(t - 1) = \alpha_{Z,t} + \tau_{Z,t} \\
    & \alpha_{X,t} | \hat{X}(t - 1), \hat{Z}_{PAR}(t - 1) \sim \text{Binomial} \left( \hat{X}(t - 1), \frac{\hat{Y}(t - 1)}{N} \right) \\
    & \alpha_{Z,t} | \hat{X}(t - 1), \hat{Z}(t - 1) \sim \text{Binomial} \left( \hat{Y}(t - 1), \gamma \right) \\
    & \tau_{X,t} \sim N \left( 0, \sigma_{\hat{X}}^{2} \right) \\
    & \tau_{Z,t} \sim N \left( 0, \sigma_{\hat{Z}}^{2} \right)
\end{align*}
\]

The process error in our simulations is shown through the Binomial draws, which noisily approximate the Kermack and McKendrick SIR equations. Measurement error is added through autoregressive error which can compound over time. In this simulation, we set \( R_0 = 2 \), which is an upper bound estimate of \( R_0 \) for pandemic influenza. We let the total population size be \( N = 100,000 \), approximately the size of an average US county. The rest of the parameter specifications for the baseline simulation data are shown in Table 2.

| Parameter | Value | 
|-----------|-------|
| \( \beta \) | 0.02 |
| \( \gamma \) | 0.01 |
| \( R_0 \) | 2.00 |
| \( T \) | 365 |
| \( X(0) \) | 99950 |
| \( Y(0) \) | 50 |
| \( Z(0) \) | 100 |
| \( \sigma_X \) | 5 |
| \( \sigma_Z \) | |

We examine three variations to the baseline data: changes in time points used \( (T) \), susceptible to initial infectious ratio \((X(0) : Y(0))\), and total population size \( N = X(0) + Y(0) + Z(0) \). Additionally, we examine two data sets where non-Kermack and McKendrick SIR models are the true generative models. The first data set is two independent SIR models and the second data set is an SEIR (XEYZ) model with \( R_0 = 2 \) such that the \( X \) and \( E \) compartments are combined together (i.e. exposed individuals are observed to be susceptible). More details of how these non-Kermack and McKendrick SIR data sets are simulated are described in Appendix A.

### 3.2 2009 Pandemic influenza data

The data for the 2009 pandemic influenza are publicly available from the Center of Disease Control and Prevention’s (CDC) FluView [FluView, 2018]. The data source is ILINet and is the national level data. We use a similar date range as in [Towers and Feng, 2009] who use data from epiweeks 21-33 (May 23, 2009-August 22, 2009). This time range turned out to be the first wave of H1N1 in the USA. The disease later had a more severe outbreak, but we do not analyze that period here.

From the United States Census Bureau, we use the population estimate of the USA as of April 1, 2010 of \( N = 308,740,000 \) individuals [United States Census Bureau, 2018]. The main features in this data set are weighted influenza like illness (wILI) for a given week (Epiweek). The wILI values are estimated based on the number of patients diagnosed with influenza like illness and weighted based on the reporting locations. To obtain weekly incidence from wILI, we use \( J(t) = \text{wILI}(t) / 100 \cdot N \).

To impute the number of individuals in each state at different time points, we use the relations described in Eq. (13), which requires an estimate of \( \gamma^{-1} \), the average expected time from infection to recovery. We use \( \gamma^{-1} = 3 \) days which was estimated by [Colizza, Barrat, Barthelemy, Valleron, and Vespignani, 2007] and was also the estimate used by [Towers and Feng, 2009]. Because \( \gamma^{-1} \) is a random variable, we perform sensitivity analysis and report \( R_0 \) estimates for multiple values of \( \gamma^{-1} \). The data is summarized in Table 3 and the pre-processed data is shown in Figure 2. Looking
Influenza: Incidence to SIR values

For different values of average time to infection $\gamma^{-1}$

![Graph showing the number of cases in states X, Y, and Z over time for different values of $\gamma^{-1}$ (days)].

Figure 2: Observed data from H1N1 influenza 2009 in the USA. Weekly wILI is reported, which we transform into S, I, and R counts shown for weeks April 16-Aug. 27. We see large variation in both the number and the shape of the Y and Z compartments for values between 2 and 10 days of the average time to recovery ($\gamma^{-1}$).

at the pre-processed data with differing values of $\gamma^{-1}$, we see that the course of the outbreak can drastically differ depending on the average infectious period.
Table 4: Table of $R_0$ point estimates and standard error (SE) for the baseline simulated data. These numbers correspond to the forest plots in Figure 3.

| Methods | $\hat{R}_0$ (SE($\hat{R}_0$)) |
|---------|-----------------------------|
| EG      | 1.002 (1e-04)               |
| RE      | 2.021 (0.0105)              |
| rRE     | 2.021 (0.0106)              |
| LL      | 2.014 (0.01)                |
| MC      | 1.31 (<1e-04)              |
| LBE     | 1.981 (0.0123)              |
| IPR     | 16.482 (21.8991)            |
| LMA     | -0.944 (0.9473)             |

Table 3: Summary of the data used to estimate $R_0$ during the 2009 H1N1 influenza pandemic in the USA. We follow the data collection and SIR imputation procedure described in Towers and Feng (2009) (although we begin our analysis on April 16 instead of May 23) and shown in Eq. (13).

| Feature           | Notes |
|-------------------|-------|
| Disease           | H1N1 Pandemic Influenza A |
| Source            | CDC FluView (https://www.cdc.gov/flu/weekly/) |
| Year              | 2009 |
| Date(s)           | April 16-August 22 (Epiweeks 15-33) |
| Location          | USA |
| Population        | N=308,740,000 people |
| Avg. Duration ($\gamma^{-1}$) | 3 days (and sensitivity analysis) |
| Raw data          | Weekly wILI reports |
| Imputed data      | $X(t), Y(t), Z(t)$ |

4 Results

4.1 Simulation Results

We present the results of a single, arbitrary simulation for each specified set of initial parameters. The reason for this is 1) because epidemic data typically occurs within a single data set and 2) so readers see how the point and CI estimates look together. As a note, we ran thousands of simulations that are not shown here but are available for reproduction in our code repository at github.com/skgallagher/r0. The results described and shown here are representative of our simulations as a whole.

We find the estimates from RE, rRE, LL, and LBE to be accurate and informative (in terms of CI size) and the estimates of EG, MC, IPR, and LMA to be both inaccurate and uninformative. Point and 95% CI estimates for the baseline simulated data are shown in Figure 3 and the corresponding table of numbers are shown in Table 4. These results indicate that both 1) a variety of methods can be used to reach similar conclusions such as RE, rRE, LL, LBE all indicate that $\hat{R}_0 \approx 2$ and 2) that naive application of methods can lead to poor interpretations of $R_0$ (EG, MC, IPR, LMA). For example, we see that exponential growth is not a good approximation of the Kermack and McKendrick SIR model and should not be used to obtain ‘back-of-the-envelope’ estimates of $R_0$.

In our simulations, we find that the unspecified-distribution methods of RE, rRE, and LL produce accurate and informative estimates of $R_0$, which may be especially useful when the specifics of disease transmission are unknown. On the other hand, our likelihood based method of LBE, a fairly simple model of disease transmission, produces comparable results to that of RE, rRE, and LL. Since the basic LBE model is an accurate and informative estimate, it can be used to verify more complex likelihood based models where more is known about the disease transmission process. Note that the successful estimates closely model the true transmission function, especially in comparison to EG, MC, and LMA which make considerable approximations to the Kermack and McKendrick SIR model. IPR is an outlier in that although it closely models the transmission function, it produces estimates that are neither reliable or informative.

In addition to showing estimates of $R_0$ for the baseline simulated data, we also show the difference among the methods when we vary three important disease parameters, the number of time steps used ($T$), the initial percent of infectious
Figure 3: Point and 95% CI estimates of a result of a single simulated data set for the given parameter conditions. The vertical blue line denotes the true value of $R_0$. The estimates colored in red have CIs that are truncated at least on one side for visibility. The corresponding table of estimates are shown in Table 4.

individuals $Y(0)/(X(0) + Y(0))$, and the population size $N = X(0) + Y(0)$. These results show variation from the baseline simulated data in one parameter at a time and are summarized visually in Figure 4 and Table 5. We only show the results from the methods of RE, rRE, LL, and LBE so as to more closely examine differences among our more successful method estimates.

With respect to the number of time points used (left column in Fig. 4), we see that confidence intervals increase as the number of time points decreases. This is especially apparent for LBE. When $T = 20$ (early in the outbreak), note how all methods overestimate the true value of $R_0$. With respect to the initial percent of infectious individuals (middle column in Fig. 4), for the most part the estimates of $R_0$ from RE, rRE, LL, and LBE are accurate and informative, except when there is exactly one infectious individual. In this case, we see RE produces a very large CI. With respect to the population size, we find that estimates decrease in accuracy and increase in CI width as $N$ decreases. This is especially apparent when the initial population size is quite small ($N=100$), where none of the estimates produce an accurate estimate of $R_0$ and only LL produces a somewhat informative estimate.

In some of the estimates from LBE in Figure 4, $\hat{R}_0$ has wider CI than we would perhaps expect or want. Part of this is that LBE is particularly sensitive to the data generation process, namely in that it requires the number of susceptibles to be monotonically non-increasing and the number of recovered to be monotonically non-decreasing. However, the autoregressive observation error we add in the simulation is not constrained to this (and we have observed data in the wild that also defies this constraint). This goes to highlight the nuances and difficulties involved in these estimates.

Finally, we show estimates and 95% CIs where the simulated data is generated from non-Kermack and McKendrick SIR models in Figure 5 and Table 6. When we misidentify XEYZ (SEIR) data as XYZ (SIR) data, then we see we systematically (but very slightly) underestimate $R_0$. While mathematically, we would expect underestimation of $R_0$, this is likely worse real life implications than overestimating $R_0$. On the other hand, if there are two separate groups of equal size that have two different true values of $R_0$ (1.5 and 2.5), the combined estimate of the population $R_0$ is about 1.75. This shows some of the non-linearity about the quantity $R_0$, as it is not as simple as taking the average $R_0$ of the equally sized groups. This may indicate that correctly identifying sub-populations that have different risk factors is important to create accurate and informative estimates of $R_0$. Overall, we find that proper model specification is very important when obtaining an estimate of $R_0$.

### 4.2 2009 Pandemic Influenza Results

The forest plot for the eight method estimates (assuming $\gamma^{-1} = 3$ days when imputing SIR data from the wILI reports) is displayed in Figure 6 and their corresponding table of values is seen in Table 7. As this is real data, we cannot say what the true value of $R_0$ is, but encouragingly, the methods that had the best results in the simulation study (RE, rRE, LL, and LBE) have similar point estimates. The median estimate is $\hat{R}_0 = 1.23$. Surprisingly, even EG and MC have estimates close to 1.20, even though they rarely matched RE, rRE, LL, or LBE in our simulation study. The full table of
Figure 4: Point and 95% CI estimates of a result of a single simulated data set for the given parameter conditions. The vertical blue line denotes the true value of $R_0$. The estimates colored in red have CIs that are truncated at least on one side for visibility. The corresponding table of estimates are shown in Table 5.

Table 5: Table of $R_0$ point estimates and standard error (SE) for simulated data that differs from the baseline simulated data by either number of time steps, population size, and initial percent of infectious. These numbers correspond to the forest plots in Figure 4.

| Condition | RE         | rRE        | LL         | LBE         |
|-----------|------------|------------|------------|-------------|
| $T = 200$ | 2.019 (0.0052) | 2.019 (0.0049) | 2.014 (0.0085) | 1.994 (0.0081) |
| $T = 100$ | 1.968 (0.0607) | 1.968 (0.0608) | 1.96 (0.0688) | 1.899 (0.0723) |
| $T = 50$  | 1.994 (0.2058) | 1.994 (0.2057) | 1.975 (0.2314) | 3.068 (2.2356) |
| $T = 20$  | 2.999 (0.7789) | 2.999 (0.7785) | 2.094 (1.7922) | 3.133 (2.4472) |
| $(X(0) = 99,999, Y(0) = 1)$ | 2.545 (3.0645) | 2.086 (0.0742) | 1.746 (0.5873) | 1.962 (0.0168) |
| $(X(0) = 99,990, Y(0) = 10)$ | 2.013 (0.0213) | 2.012 (0.0212) | 2.006 (0.0124) | 1.961 (0.0148) |
| $(X(0) = 99,900, Y(0) = 100)$ | 2.006 (0.0117) | 2.006 (0.0118) | 2.003 (0.0117) | 1.969 (0.0146) |
| $(X(0) = 99,000, Y(0) = 1,000)$ | 2.008 (0.0131) | 2.008 (0.0132) | 2.007 (0.0126) | 1.962 (0.0186) |
| $(X(0) = 999,500, Y(0) = 500)$ | 2.015 (0.0038) | 2.015 (0.0035) | 2.011 (0.0032) | 2.000 (0.0028) |
| $(X(0) = 9,950, Y(0) = 50)$ | 2.032 (0.0221) | 2.032 (0.0222) | 2.025 (0.025) | 1.283 (0.2032) |
| $(X(0) = 990, Y(0) = 10)$ | 2.014 (0.1288) | 2.014 (0.1289) | 2.494 (0.3402) | 0.054 (1.9697) |
| $(X(0) = 99, Y(0) = 1)$ | -55 (1230) | 21 (41) | 1.471 (0.4831) | 2.168 (4.6) |
Figure 5: Point and 95% CI estimates of a result of a single simulated data set for the given parameter conditions. The vertical blue line denotes the 'true' value of $R_0$. The estimates colored in red have CIs that are truncated at least on one side for visibility. The corresponding table of estimates are shown in Table 6. The figure entitled “SEIR” refers to data generated according to Eq. (16), and the data in “Sum SIR” refers to two SIR models that have been combined together, one with a true value of $R_0 = 1.5$ and the other with $R_0 = 2.5$. These forest plots correspond to the figures in Table 5.

Table 6: Table of $R_0$ point estimates and standard error (SE) for non-Kermack and McKendrick SIR, simulated data. These numbers correspond to the forest plots in Figure 6.

| Methods | SEIR $\hat{R}_0$ ($SE(\hat{R}_0)$) | SIR $\times 2$ $\hat{R}_0$ ($SE(\hat{R}_0)$) |
|---------|---------------------------------|---------------------------------|
| EG      | 1.001 ($<1e-04$)               | 1.002 (2e-04)                   |
| RE      | 1.977 (0.0161)                 | 1.767 (0.137)                   |
| rRE     | 1.977 (0.0162)                 | 1.767 (0.0795)                  |
| LL      | 1.989 (0.021)                  | 1.789 (0.0042)                  |
| MC      | 1.31 (0)                      | 1.619 (0)                       |
| LBE     | 1.893 (0.0135)                 | 1.742 (5.5501)                  |
| IPR     | 4.254 (0.8687)                 | 1.177 (0.9889)                  |
| LMA     | -0.645 (2.5491)               | 2.565 (0.175)                   |
Figure 6: Results from SIR methods on H1N1 data, assuming $\gamma^{-1} = 3$ days when imputing SIR data from the wILI reports.

results is shown in Table 7. Using the results from rRE, we estimate that $\hat{R}_0 = 1.23$ with 95% CI [1.03, 1.42], which is a smaller estimate than the value of 1.7-1.9 found in Towers and Feng (2009), which used the same estimate of $\gamma^{-1} = 3$ days.

Table 7: Estimates of $R_0$ from the 8 methods and assuming $\gamma^{-1} = 3$ when imputing SIR data from the wILI reports.

| Method | $\hat{R}_0$ | $SE(\hat{R}_0)$ |
|--------|-------------|-----------------|
| EG     | 1.220       | 0.0899          |
| RE     | 1.234       | 0.3281          |
| rRE    | 1.234       | 0.0959          |
| LL     | 1.214       | 0.0214          |
| MC     | 1.141       | 0.0000          |
| LBE    | 1.345       | 3.4319          |
| IPR    | 1.293       | 0.8059          |
| LMA    | -1.552      | 1.0699          |

In Figure 7, we show the number of weeks used to estimate $R_0$, all with the initial time point of April 16, 2019 and using one of 4, 8, 12, or 16 weeks worth of data. We see that the average estimate for most of the $R_0$ estimates are robust to the amount of time used but the estimates of $R_0$ from 4 weeks of data compared to the others, results in larger estimates of $R_0$. The method estimates are reported in Table 8. When using only 4 weeks of data, we see estimates of $R_0$ from LL to be 1.80 (0.33), which is much closer to the Towers and Feng (2009) estimate.

Our above results are generated after imputing the SIR data from wILI counts using the average time to recovery, $\gamma^{-1} = 3$ days, which is a major assumption. In Figure 8 it is clear that estimates of $R_0$ increase in magnitude as the estimate of $\gamma^{-1}$ increases, which makes sense as a longer infectious period leads to more infections. Moreover, we see that the width of the CIs typically increases for the estimates as $\gamma^{-1}$ increases and is especially noticeable for rRE. On the other hand, we see that the LBE CI widths are large, which may indicate that the transmission spread does not follow the simple specified likelihood model well. Finally, we see that both EG and MC produce estimates that only increase by a few thousandths for different periods of infectiousness. In contrast, the mean estimate from RE is 1.23 for $\gamma^{-1} = 3$ days and is 1.87 when $\gamma^{-1} = 9$ days. In context, the differences in estimates represent a final size (cumulative percent of infected individuals) difference from 35% to 76%, respectively, or approximately 123 million people in the US.

That said, all of RE, rRE, LBE, and LL have similar point estimates of $R_0$ for a given value of $\gamma^{-1}$ in the imputation step. We also note that in all their estimates, RE and LBE have SE of at least 0.25, which means the confidence intervals
Figure 7: Results from SIR methods on H1N1 data, assuming $\gamma^{-1} = 3$ days when imputing SIR data from the wILI reports. This shows the estimates of $R_0$ using different maximum time points, to get a sense of how $R_0$ estimates can change with time. The top left graph uses 4 weeks, the top right 8 weeks, bottom left 12 weeks, and bottom right 16 weeks worth of data. The corresponding table of estimates is shown in Table 8.

Table 8: Results from SIR methods on H1N1 data, assuming $\gamma^{-1} = 3$ days when imputing SIR data from the wILI reports. This shows the estimates of $R_0$ using different maximum time points, to get a sense of how $R_0$ estimates can change with time. The table corresponds to the forest plots shown in Figure 7.

| Method | $\hat{R}_0$–4 weeks | $\hat{R}_0$–8 weeks | $\hat{R}_0$–12 weeks | $\hat{R}_0$–16 weeks |
|--------|---------------------|---------------------|----------------------|----------------------|
| EG     | 1.336 (0.1678)      | 1.167 (0.0592)      | 1.117 (0.0339)       | 1.098 (0.0245)       |
| RE     | 2.179 (0.2689)      | 1.532 (0.2815)      | 1.341 (0.3264)       | 1.247 (0.3516)       |
| rRE    | 2.176 (0.5043)      | 1.53 (0.2671)       | 1.341 (0.1879)       | 1.247 (0.1254)       |
| LL     | 1.796 (0.33)        | 1.329 (0.0802)      | 1.236 (0.0374)       | 1.193 (0.0232)       |
| MC     | 1.141 (<1e-04)      | 1.141 (<1e-04)      | 1.141 (<1e-04)       | 1.141 (<1e-04)       |
| LBE    | 1.885 (0.3735)      | 1.427 (0.3735)      | 1.317 (0.3735)       | 1.276 (0.3735)       |
| IPR    | 2.376 (1.272)       | 1.7 (1.1037)        | 1.437 (0.9633)       | 1.29 (0.8661)        |
| LMA    | -0.983 (8.1365)     | 0.08 (4.5233)       | 0.149 (3.4279)       | -1.932 (6.9396)      |

are fairly large. Of the methods we prefer, only LL has a reasonably small CI width for every value of $\gamma^{-1}$, with a maximum SE of 0.06.
Figure 8: Variability of $R_0$ estimates for different values of $\gamma^{-1}$, the average time to recovery or average infectious period in imputing the SIR data from the wILI reports.

Overall, from this analysis, we conclude when we assume that $\gamma^{-1} = 3$ days, that $R_0$ is closer to 1.2 than it is to 1.6, which indicates that, at least within this time frame, $R_0$ for H1N1 is comparable to that of seasonal influenza where $\hat{R}_0$ is estimated to be between 1.19-1.37 [Biggerstaff, Cauchemez, Reed, Gambhir, & Finelli, 2014]. However, our estimate does not factor in how prevention/awareness campaigns influenced the spread of the disease, the differing seasonality of pandemic influenza (the time range we analyze is May-August) compared to seasonal influenza (the time range is typically October-March), and the severity and impact of the infection on the individual. Another possible concern is that of “backfill,” the phenomenon of where the number of infections for the different weeks are constantly updated, even months after the original date. As such, we have no guarantee we are using the same numbers from previous analyses. Also, pandemic influenza occurred in two waves in the US, and we are analyzing the first wave, as was done in [Towers and Feng, 2009].

We assumed that $\gamma^{-1} = 3$ days when imputing the SIR data from wILI reports. If we had assumed a 7 or 9 day infectious period then our estimates of $R_0$ become more in line with those reported by Biggerstaff et al. (2014) for pandemic influenza: estimates of $\hat{R}_0$ between 1.3-1.7. Finally, we see that the estimates of $R_0$ are sensitive to the time period used (see Table [8]) where 4 weeks of data lead to estimates of about 1.8 or larger and which substantially decreases as more data points (time) is added.

What this disparity in estimates indicates is how sensitive estimates of $R_0$ can be to different parameters and emphasizes the need for reproducible research in this area. The reproduction number can be a viable way to compare disease outbreaks to one another, but it is not so simple as comparing one number to another, as the model framework (e.g. SIR vs. SEIR) must be considered. We recommend, along with reporting the estimate and CI of $R_0$ to also explicitly report the modeling framework (e.g. Transmission Model: $S \rightarrow I \rightarrow R$, $\hat{R}_0 = 1.4$, 95% CI [1.1, 1.5]) and what steps are taken in pre-processing the data and what time steps are used and why.

5 Discussion

In this paper, we examine the nuances of the initial reproduction number $R_0$, with emphasis on 1) variability in methods of estimation, 2) sensitivity to a number of infection parameters, and 3) the role of data pre-processing steps to transform incidence counts to a more familiar data format.

As to issue 1), we examine eight methods to estimate $R_0$ which include exponential growth (EG), ratio estimator (RE), reparameterized ratio estimator (rRE), log linear (LL), Markov Chain (MC), likelihood-based estimate (LBE), incidence-to-prevalence ratio (IPR), and linear model approximation (LMA). Although, this is not an exhaustive list of methods used to estimate $R_0$ from the SIR model, we believe it shows the diversity of the different methods that may be used to estimate this quantity. We review the details of these methods in Sections 2.1, 2.2, 2.8

As to issue 2), we compare the the estimates of $R_0$ for the eight methods using a series of simulations. To this effort, we generate different simulated data sets under different noise assumptions and estimate $R_0$ using each of these data sets. We then analyze estimates of $R_0$ while varying individual parameters of the model including the infection and recovery
parameters, the total amount of data, the initial percent of infectious and susceptible individuals, the standard deviation of the noise used to generate the simulated data, and the total number of individuals in a population. We also analyze the model under misspecification, that is when a model other than the SIR model is used to generate the simulated data sets. Our simulation code is available for reproduction in our code repository: [github.com/skgallagher/r0](https://github.com/skgallagher/r0).

As to issue 3) we discuss one common approach to pre-processing incidence counts into the number of susceptible, number of infectious, and number of recovered individuals at each time step and how this pre-processing is dependent on our estimate of the average recovery period duration, $\gamma^{-1}$. We show the effects of this pre-processing step in our application to 2009 pandemic influenza in the US.

Overall, we recommend based on our analysis:

- **RE, rRE, LL, and LBE** provide the most accurate and informative estimates of $R_0$. This is based on both the results of our simulation study and the similar results in the 2009 pandemic influenza data application.
- **LL** provides a way to establish a useful novel/secondary/confirmatory interpretation of $R_0$.
- **LBE** is a reliable and useful estimate and tends to generate larger CIs than the sum of square based methods (RE, rRE, LL).
- **LBE** is a simple likelihood based estimate and seems to work well in our simulation study and the data application, which shows to us that other more sophisticated LBE methods may work even better.
- We do not recommend using EG or IPR even as ‘back of the envelope’ calculations based on our simulation study.
- The larger the number of initially infectious individuals is, the smaller the CIs for $R_0$ are.
- Of the parameters described here, $\hat{R}_0$ is most sensitive to changes in $T$, the time points used in an epidemic.
- Small sample sizes in total population size or the initial percent of infectious also cause disparity in $R_0$ method estimates and larger CIs than when changing other simulation parameters.

We also analyzed estimates of $R_0$ using data from the 2009 H1N1 pandemic in the USA. From this study we conclude:

- For the H1N1 application, we find that when we assumed the infectious period was on average $\gamma^{-1} = 3$ days, we found our estimates to be closer to $\hat{R}_0 = 1.2$, whereas the [Towers and Feng (2009)](https://journals.ametsoc.org/view/journals/m Chern/1/1/1557168x_a1b80077.full.html) reports estimates of 1.7-1.9 from. In both our and their study, we use similar data imputation steps, most importantly assuming $\gamma^{-1}=3$ days.
- Also for H1N1, we found that when we increased $\gamma^{-1}$ to 7 or 9 days, we find estimates much more in line with 2009 pandemic influenza values of $R_0$ reported in [Biggerstaff et al. (2014)](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4027637/).
- From this, we conclude that imputation/pre-processing steps are essential in producing accurate and reliable estimates of $R_0$.

Overall, we find that $R_0$ is a difficult quantity to estimate. We find that estimates often are nuanced and only interpretable in the context of other parameters, which is contrary to the idea of a one number summary, which $R_0$ is often purported to be. We, however, are not yet ready to give up on the ‘most important quantity’ in epidemic modeling.

To better understand what the process of estimating $R_0$ entails, we recommend the following guidelines. First, we emphasize the need for reproducible research in this field, in every step there is: data collection, data cleaning, and statistical analysis. We also recommend that every abstract/paper summary not only includes the estimate and 95% CI of $R_0$ but also explicitly gives the disease framework and any pre-processing/imputation steps of the data. This would go a long way for scientists to methodically scrape publication data to create an extensive table of $R_0$ for various outbreaks around the world. Finally, we do not see any harm in presenting multiple estimates of $R_0$, especially those of rRE, RE, LL, and LBE. Disparity in these estimates, for example, can show a deeper issue about the data.

The goal of this paper is explore some of the underlying nuances of $R_0$ how these nuances shape and guide ongoing statistical analysis. We hope this paper provides a step in the right direction in a better understanding of $R_0$ for all scientists involved in the field of infectious disease transmission.

As this paper is being prepared, COVID-19, novel coronavirus, is beginning to severely outbreak in the US, and we see many estimates of $R_0$ being purported online and in the news (such as a table seen in [MIDAS (2020)](https://www.cdc.gov/coronavirus/2019-ncov/disease-severity/hospitalization.html)). We hope this paper provides guidance in what features to analyze and assess when attempting to assess the reliability of $R_0$ estimates.
References

Abbey, H. (1952). An examination of the Reed-Frost theory of epidemics. *Human biology, 24*(3), 201.

Anderson, R., & May, R. (1992). *Infectious diseases of humans*. Oxford: Oxford University Press.

Bhadra, A., Ionides, E. L., Laneri, K., Pascual, M., Bouma, M., & Dhaman, R. C. (2011). Malaria in northwest India: Data analysis via partially observed stochastic differential equation models driven by levy noise. *Journal of the American Statistical Association, 106*(494), 440-451. Retrieved from [https://doi.org/10.1198/jasa.2011.ap10323](https://doi.org/10.1198/jasa.2011.ap10323).

Biggerstaff, M., Cauchemez, S., Reed, C., Gambhir, M., & Finelli, L. (2014). Estimates of the reproduction number for seasonal, pandemic, and zoonotic influenza: a systematic review of the literature. *BMC infectious diseases, 14*(1), 480.

Blower, S. M., & Chou, T. (2004). Modeling the emergence of the ‘hot zones’: tuberculosis and the amplification dynamics of drug resistance. *Nature medicine, 10*(10), 1111–1116.

Britt, T. (2010). Stochastic epidemic models: A survey. *Mathematical Biosciences, 225*(1), 24-35. Retrieved from [http://www.sciencedirect.com/science/article/pii/S0025556410000143](http://www.sciencedirect.com/science/article/pii/S0025556410000143).

Brown, G. D., Oleson, J. J., & Porter, A. T. (2016). An empirically adjusted approach to reproductive number estimation for stochastic compartmental models: A case study of two ebola outbreaks. *Biometrics, 72*(2), 335–343.

Chen, G., & Li, T. (2009). Stability of stochastic delayed sir model. *Stochastics and Dynamics, 9*(2), 231–252.

Colizza, V., Barrat, A., Barthelemy, M., Valleron, A.-J., & Vespignani, A. (2007, 01). Modeling the worldwide spread of pandemic influenza: Baseline case and containment interventions. *PLOS Medicine, 4*(1), 1-16. Retrieved from [https://doi.org/10.1371/journal.pmed.0040013](https://doi.org/10.1371/journal.pmed.0040013).

Craven McGinty, J. (2020, Feb). How many people might one person with coronavirus infect? *Wall Street Journal*. Retrieved from [https://www.wsj.com/articles/how-many-people-might-one-person-with-coronavirus-infect-11581676200](https://www.wsj.com/articles/how-many-people-might-one-person-with-coronavirus-infect-11581676200).

Daley, D. J., Gani, J., & Gani, J. M. (2001). *Epidemic modelling: an introduction*. Cambridge University Press.

Diekmann, O., Heesterbeek, J., & Roberts, M. (2009). The construction of next-generation matrices for compartmental epidemic models. *Journal of the Royal Society Interface, rsif20090386. doi: 10.1098/rsif.2009.0386. 2009*(10), 1-16. Retrieved from [https://doi.org/10.1098/rsif.2009.0386](https://doi.org/10.1098/rsif.2009.0386).

Epstein, J. M. (2007). Agent-based computational models and generative social science. *Introductory Chapters. doi: 10.1093/biostatistics/4.2.279*. Retrieved from [https://doi.org/10.1093/biostatistics/4.2.279](https://doi.org/10.1093/biostatistics/4.2.279).

Farrington, C. P., Kanaan, M. N., & Gay, N. J. (2003, 04). Branching process models for surveillance of infectious diseases controlled by mass vaccination. *Biostatistics, 4*(2), 279-295. Retrieved from [https://doi.org/10.1093/biostatistics/4.2.279](https://doi.org/10.1093/biostatistics/4.2.279).

Fintzi, J., Cui, X., Wakefield, J., & Minin, V. N. (2017). Efficient data augmentation for fitting stochastic epidemic models to prevalence data. *Journal of Computational and Graphical Statistics, 26*(4), 918-929. Retrieved from [https://doi.org/10.1080/10618600.2017.1328365](https://doi.org/10.1080/10618600.2017.1328365).

FluView, C. (2018). 2008-2009 influenza season. Available at [https://gis.cdc.gov/grasp/fluview/fluportaldashboard.html](https://gis.cdc.gov/grasp/fluview/fluportaldashboard.html). Accessed June, 18, 2018.

Forsberg White, L., & 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*(16), 2999–3016. Retrieved from [http://dx.doi.org/10.1002/sim.3136](http://dx.doi.org/10.1002/sim.3136).

Gani, J., & Yakowitz, S. (1995). Error bounds for deterministic approximations to markov processes, with applications to epidemic models. *Journal of Applied Probability, 32*(4), 1063–1076. doi: 10.1017/S0021900200103547.

Harko, T., Lobo, F. S., & Mak, M. (2014). Exact analytical solutions of the susceptible-infected-recovered (sir) epidemic model and of the sir model with equal death and birth rates. *Applied Mathematics and Computation, 236, 184–194.

He, D., Ionides, E. L., & King, A. A. (2009). Plug-and-play inference for disease dynamics: measles in large and small populations as a case study. *Journal of The Royal Society Interface*. Retrieved from [http://rsif.royalsocietypublishing.org/content/early/2009/06/10/rsif.2009.0151](http://rsif.royalsocietypublishing.org/content/early/2009/06/10/rsif.2009.0151). doi: 10.1098/rsif.2009.0151.

Hethcote, H. W. (2000). The mathematics of infectious diseases. *SIAM Review, 42*(4), 599-653. Retrieved from [http://www.jstor.org/stable/2653135](http://www.jstor.org/stable/2653135).

Hu, Z., Teng, Z., & Long, Z. (2014). Stability and bifurcation analysis in a discrete sir epidemic model. *Mathematics and computers in Simulation, 97*, 80–93.
Jacquez, J. A., & O’Neill, P. (1991). Reproduction numbers and thresholds in stochastic epidemic models i. homogeneous populations. *Mathematical Biosciences, 107*(2), 161 - 186. Retrieved from http://www.sciencedirect.com/science/article/pii/0025556491900032 doi: http://dx.doi.org/10.1016/0025-5564(91)90003-2

Kermack, W. O., & McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, 115*(772), 700–721. Retrieved from http://rspa.royalsocietypublishing.org/content/115/772/700 doi: 10.1098/rspa.1927.0118

King, A., Nguyen, D., & Ionides, E. (2016). Statistical inference for partially observed markov processes via the r package pomp. *Journal of Statistical Software, Articles, 69*(12), 1–43. Retrieved from https://www.jstatsoft.org/v69i12 doi: 10.18637/jss.v069.i12

Li, J., Blakeley, D., et al. (2011). The Failure of R0. *Computational and mathematical methods in medicine, 2011.* Lotka, A. J. (1920). Undamped oscillations derived from the law of mass action. *Journal of the american chemical society, 42*(8), 1595–1599.

Majumder, M. S., Santillana, M., Mekaru, S. R., McGinnis, D. P., Khan, K., & Brownstein, J. S. (2016). Utilizing nontraditional data sources for near real-time estimation of transmission dynamics during the 2015-2016 colombian zika virus disease outbreak. *JMI R public health and surveillance, 2*(1), e30. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4909981/

MIDAS. (2020). *Covid-19.* https://github.com/midas-network/COVID-19/tree/master/parameter_estimates/2019 novel coronavirus GitHub

Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *The Computer Journal, 7*(4), 308-313. Retrieved from http://dx.doi.org/10.1093/comjnl/7.4.308 doi: 10.1093/comjnl/7.4.308

Nishiura, H., & Chowell, G. (2009). The effective reproduction number as a prelude to statistical estimation of time-dependent epidemic trends. In G. Chowell, J. M. Hyman, L. M. A. Bettencourt, & C. Castillo-Chávez (Eds.), *Mathematical and statistical estimation approaches in epidemiology* (pp. 103–121). Dordrecht: Springer Netherlands. Retrieved from http://dx.doi.org/10.1097/978-90481-2313-1\_5 doi: 10.1007/978-90481-2313-1\_5

Nishiura, H., Chowell, G., Safran, M., & 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.

Obadia, T., Haneef, R., & Boëlle, P.-Y. (2012). The r0 package: a toolbox to estimate reproduction numbers for epidemic outbreaks. *BMC medical informatics and decision making, 12*(1), 147.

Towers, S., & Feng, Z. (2009). Pandemic h1n1 influenza: Predicting the course of vaccination programme in the united states. *Euro. Surveill., 14.*

United States Census Bureau. (2018). Quickfacts. *Available at: Accessed June, 18, 2018.*

Van den Driessche, P. (2017). Reproduction numbers of infectious disease models. *Infectious Disease Modelling, 2*(3), 288–303.

Wallinga, J., & Lipsitch, M. (2007). How generation intervals shape the relationship between growth rates and reproductive numbers. *Proceedings of the Royal Society of London B: Biological Sciences, 274*(1609), 599–604.

Wang, H., Wang, Z., Dong, Y., Chang, R., Xu, C., Yu, X., … others (2020). Phase-adjusted estimation of the number of coronavirus disease 2019 cases in wuhan, china. *Cell Discovery, 6*(1), 1–8.

Wasserman, L. (Ed.). (2004). *All of Statistics: A Concise Course in Statistical Inference.* Springer.

**A Non Kermack and McKendrick SIR-Based Data**

**XYZ (SIR) × 2** In this model, we have two independent XYZ models that are then aggregated together. Specifically, both models are given from the equations in Eq. (3). The first one, denoted A, is given by \(X_A(0) = 49, 950, Y_A(0) = 50, Z_A(0) = 0\) and disease parameters such that \(R_0^A = 1.5 (\beta_A = 0.015, \gamma_A = 0.01)\). The second one is denoted by B and is given by \(X_B(0) = 49, 950, Y_B(0) = 50, Z_B(0) = 0\) and disease parameters such that \(R_0^B = 2 (\beta_A = 0.02, \gamma_A = 0.01)\). We then add the two sets of compartments together:

\[
\begin{align*}
(X(t) &= X_A(t) + X_B(t), Y(t) = Y_A(t) + Y_B(t), Z(t) = Z_A(t) + Z_B(t)).
\end{align*}
\]

Then following the equations in Eq. (14) \(f = X\) and \(g = Z\).

**XEYZ (SEIR).** The deterministic SEIR (XEYZ) model (seen in Cintrón-Arias, Castillo-Chávez, Bettencourt, Lloyd, and Banks (2009)) is based upon the following ODEs:
\[
\begin{align*}
\frac{dX}{dt} &= -\frac{\beta XY}{N} \\
\frac{dE}{dt} &= \frac{\beta XY}{N} - \mu E \\
\frac{dY}{dt} &= \mu E - \gamma Y \\
\frac{dZ}{dt} &= \gamma Y,
\end{align*}
\]
(16)

again, with a constant population of \( N \) and with known initial values \((X(0), E(0), Y(0), Z(t))\). As in the XYZ model, \( \beta \) and \( \gamma \) represent the average rate of infection and average rate of recovery, respectively. The parameter \( \mu \) is the rate at which latently exposed individuals become infectious. The reproduction number for the XEYZ model, is \( R_0 = \frac{\beta}{\gamma} \), like in the XYZ model.

To simulate XYZ data from the XEYZ model, we first determine the XEYZ values for a given set of initial values. Then we combine the susceptible and exposed states into a new susceptible state, \( X_{\text{new}}(t) = X(t) + E(t) \). Then following the equations in Eq. (14) \( f = X_{\text{new}} \) and \( g = Z \). In the simulations we let \( \mu = 0.1 \).