Summer 8-11-2020

Comparative Assessment of Epidemiological Models for Analyzing and Forecasting Infectious Disease Outbreaks

Kimberlyn Roosa

Computational assessment of epidemiological models for analyzing and forecasting infectious disease outbreaks

By

Kimberlyn Roosa

Mathematical modeling offers a quantitative framework for analyzing mechanisms underlying infectious disease transmission and explaining patterns in epidemiological data. Models are also commonly applied in outbreak investigations for assessing intervention and control strategies and generating epidemic forecasts in real time. However, successful application of mathematical models depends on the ability to reliably estimate key transmission and severity parameters, which are critical for guiding public health interventions.

Overall, the three studies presented provide a thorough guide for assessing and utilizing mathematical models for describing infectious disease outbreak trends. In the first study, we describe the process for analyzing identifiability of parameters of interest in mechanistic disease transmission models. In the second study, we expand this idea to simple phenomenological models and explore the idea of overdispersion in the data and how to determine an appropriate error structure within the analyses. In the third study, we use previously validated phenomenological models to generate short-term forecasts of the ongoing COVID-19 pandemic.

During infectious disease epidemics, public health authorities rely on modeling results to inform intervention decisions and resource allocation. Therefore, we highlight the importance of interpreting modeling results with caution, particularly regarding theoretical aspects of mathematical models and parameter estimation methods. Further, results from modeling studies should be presented with quantified uncertainty and interpreted in terms of the assumptions and limitations of the model, methods, and data used. The methodology presented in this dissertation provides a thorough guide for conducting model-based inferences and presenting the uncertainty associated with parameter estimation results.

Computational assessment of epidemiological models for analyzing and forecasting infectious disease outbreaks

by

Kimberlyn Roosa

BS, University of South Carolina

MPH, Georgia State University

A Dissertation Submitted to the Graduate Faculty
of Georgia State University in Partial Fulfillment
of the
Requirements for the Degree

DOCTOR OF PHILOSOPHY IN PUBLIC HEALTH

ATLANTA, GEORGIA
30303

APPROVAL PAGE

Computational assessment of epidemiological models for analyzing and forecasting infectious disease outbreaks

by

Kimberlyn Roosa

Approved:

Gerardo Chowell
Committee Chair

Ruiyan Luo
Committee Member

Richard Rothenberg
Committee Member

James Mac Hyman
Committee Member

July 27, 2020
Date

Acknowledgments

My coworkers at home: Oscar, Quasimodo, and Samus

Author’s Statement Page

In presenting this dissertation as a partial fulfillment of the requirements for an advanced degree from Georgia State University, I agree that the Library of the University shall make it available for inspection and circulation in accordance with its regulations governing materials of this type. I agree that permission to quote from, to copy from, or to publish this dissertation may be granted by the author or, in his/her absence, by the professor under whose direction it was written, or in his/her absence, by the Associate Dean, School of Public Health. Such quoting, copying, or publishing must be solely for scholarly purposes and will not involve potential financial gain. It is understood that any copying from or publication of this dissertation which involves potential financial gain will not be allowed without written permission of the author.

Kimberlyn Roosa
Signature of Author

Table of Contents

Appendices:

Chapter 1: Introduction

Mathematical modeling offers a quantitative framework for analyzing mechanisms underlying infectious disease transmission and explaining patterns in epidemiological data (1, 2). Models are also commonly used by public health researchers during outbreak investigations for assessing intervention and control strategies and generating epidemic forecasts in real time. However, the successful application of mathematical models to epidemiologic studies depends upon our ability to reliably estimate key transmission and severity parameters that are critical for guiding public health interventions.

Parameter estimates for a given system are subject to uncertainty from noise in the data and assumptions built in the model, and ignoring this uncertainty can result in misleading inferences and potentially incorrect public health policy decisions (3). A crucial aspect of epidemiological models is the identifiability of the parameters, or whether a set of parameters can be uniquely estimated from a given model and data set (4). Lack of identifiability, or non-identifiability, results in a wide range of parameter values that yield nearly equivalent fits to the data. Non-identifiability can stem from the model structure (structural identifiability) or the lack of information in the available data (practical identifiability). Practical identifiability is related to characteristics in the data, including the number of observations, the temporal resolution, or observation error.

The mathematical modeling toolkit includes phenomenological models that assess features in epidemic trajectories as well as mechanistic models that evaluate the effects of interventions or other factors on transmission dynamics. Dynamic models based on differential equations are often calibrated to infectious disease outbreak data, which typically represent a time series of new cases, where a *case* corresponds to an observable event. Further, the reported dataset corresponds to only one realization of a stochastic process, and generating more data realizations in a carefully controlled environment is not feasible in the context of real outbreaks occurring in natural environments.

For the first study, we present and illustrate a simple computational method for assessing parameter identifiability in compartmental (mechanistic) epidemic models (5). We describe a parametric bootstrap approach to generate simulated data from dynamical systems to quantify parameter uncertainty and identifiability. To demonstrate this approach, we begin with a low-complexity SEIR model and work through examples of increasingly more complex compartmental models that correspond with applications to pandemic influenza, Ebola, and Zika.

When calibrating models to data via some fitting process, the model solution for a given set of parameter values and initial conditions is typically considered to be the "mean" solution, which is then embedded into a counting process characterized by a statistical model. For example, in the first study we assume the Poisson distribution as the error structure for the parametric bootstrapping analyses, as the Poisson distribution is commonly assumed for count data. In this inference framework, the *equidispersion* property of the Poisson distribution (where the mean is equal to the variance) simplifies the inference process, limits the number of degrees of freedom, and indirectly reduces potential issues of parameter non-identifiability (6). However, empirical data may exhibit greater variability than expected based on a given statistical model.

Greater variability could point to model misspecification, such as missing crucial information about the epidemiology of the disease or changes in population behavior. Hence, researchers could fix this lack of model fit by identifying and incorporating key process components in the model, thus resolving the apparent overdispersion issue. Therefore, identifying the relevant sources of apparent overdispersion is critical in the modeling process as it could lead to poor descriptions of the data and predictive power and underestimated standard errors and confidence intervals (7). When the mechanism producing the apparent overdispersion is unknown, however, it is typically assumed that the variance in the data exceeds the mean by some factor. In this case, the researcher may reconsider the error structure to allow for the variance to be larger than the mean and better represent the data (e.g., negative binomial) (6).

Simulation studies can be useful for evaluating the impact of various forms of misspecification when calibrating a model to data. In the second study, we evaluate the effects of misspecifying the error structure on the bias and uncertainty of parameter estimates for simple dynamic transmission models (8). Specifically, we focus on modeling varying levels of data overdispersion stemming from randomness in the counting process that shapes the time series data, rather than systematic misspecifications in the mean process linked to the model. For this study, we analyze two phenomenological models – the generalized growth model and generalized logistic growth model – to assess how results of parameter estimation are affected by the level of overdispersion in the data. We utilize the parametric bootstrap approach described in the first study to assess parameter estimates and their uncertainty as a function of the level of random noise in the data, and we compare results using two common parameter estimation methods: nonlinear least squares (LSQ) and maximum likelihood estimation with a Poisson error structure (Poisson-MLE).

Phenomenological growth models, like those presented in the second study, that are able to capture the empirical patterns of past epidemics can be used to investigate the trajectory of epidemics in real time and are especially useful when the amount of epidemiological data are limited (3, 9-11). Real-time short-term forecasts generated from such models can be useful to allocate the resources needed to bring an epidemic under control. The ongoing COVID-19 pandemic, for example, originated in December 2019 in Wuhan, China, where cases quickly outnumbered the available number of beds in hospitals, putting a substantial burden on the healthcare system. To anticipate additional resources needed to combat the epidemic, mathematical and statistical modeling tools were used to generate timely short-term forecasts of reported cases and estimates of expected morbidity burden that can help guide public health preparation. Short-term forecasts can also guide the intensity and type of interventions needed to mitigate an epidemic.

For the third study, we employ previously validated phenomenological models, including the generalized logistic growth model presented in the second study, and apply to the ongoing COVID-19 pandemic to generate short-term forecasts for two provinces in China (12). These forecasts were generated in February 2020, when the epidemiological features of COVID-19 were still unclear, and the disease continued to spread within and outside of China, despite several social distancing measures implemented by the Chinese government. Limited epidemiological data were available, and recent changes in case definition and reporting further complicated our understanding of the impact of the epidemic. All of these factors obscured the true underlying epidemic trajectory and complicated inference of epidemiological parameters and the calibration of mechanistic transmission models. Therefore, we employ these phenomenological models to generate short-term forecasts of the cumulative reported cases of COVID-19 in Guangdong and Zhejiang, China.

Overall, the three studies in Chapters 2 – 4 provide a thorough guide for assessing and utilizing mathematical models for describing infectious disease outbreak trends. Chapter 2 describes the process for analyzing identifiability of parameters of interest in a given disease transmission model, specifically focusing on mechanistic models ranging from simple to complex. Chapter 3 expands this idea to simple phenomenological models and explores the idea of overdispersion and how to determine an appropriate error structure within the analyses. Chapter 4 utilizes previously validated phenomenological models, including one presented in Chapter 3, to generate short-term forecasts of the ongoing COVID-19 pandemic, where public health authorities have relied on modeling results to inform intervention decisions and resource allocation.

Chapter 2. Assessing parameter identifiability in compartmental dynamic models using a computational approach: application to infectious disease transmission models

Roosa, K. & Chowell, G. (2019). Assessing parameter identifiability in compartmental dynamic models using a computational approach: application to infectious disease transmission models. *Theoretical Biology and Medical Modeling*. doi:10.1186/s12976-018-0097-6

1. Introduction

Mathematical modeling is commonly applied in outbreak investigations for analyzing mechanisms behind infectious disease transmission and explaining patterns in epidemiological data (1, 2). Models also provide a quantitative framework for assessing intervention and control strategies and generating epidemic forecasts in real time. However, the successful application of mathematical modeling to investigate epidemics depends upon our ability to reliably estimate key transmission and severity parameters, which are critical for guiding public health interventions. In particular, parameter estimates for a given system are subject to two major sources of uncertainty: noise in the data and assumptions built in the model (3). Ignoring this uncertainty can result in misleading inferences and potentially incorrect public health policy decisions.

Appropriate and flexible approaches for estimating parameters from data, evaluating parameter and model uncertainty, and assessing goodness of fit are gaining increasing attention (4-8). For instance, model parameters can be estimated by connecting models with observed data through various methods, including least-squares fitting (9), maximum likelihood estimation (10, 11), and approximate Bayesian computation (12, 13). An important, yet often overlooked step in estimating parameters is examining parameter identifiability – whether a set of parameters can be uniquely estimated from a given model and data set (14). Lack of identifiability, or non-identifiability, occurs when multiple sets of parameter values yield a very similar model fit to the data. Non-identifiability may be attributed to the model structure (structural identifiability) or due to the lack of information in a given data set (practical identifiability), which could be associated with the number of observations, spatial-temporal resolution (e.g., daily versus weekly data), and observation error. A parameter set is considered structurally identifiable if any set of parameter values can be uniquely mapped to a model output (15). As such, structural identifiability is the first step in understanding which model parameters can be estimated from data of certain state(s) of the system at a specific spatial-temporal resolution. Structurally identifiable parameters may still be non-identifiable in practice due to a lack of information in available data. The so-called “practical identifiability” considers real-world data issues: amount of noise in the data and sampling frequency (e.g., data collection process) (14).

Several methods have been proposed to examine structural identifiability of a model without the need of experimental data; these include Taylor series methods (15, 16), differential algebra-based methods (17, 18), and other mathematical approaches (15, 19). These methods tend to work better in the context of simple rather than complex models. Model complexity, in general, is a function of the number of parameters necessary to characterize the states of the system and the spectrum of dynamics that can be recovered from the model. Model complexity affects the ability to reliably parameterize the model given the available data (3), so there is a need for flexible, mathematically-sound approaches to address parameter identifiability in models of varying complexity. Here, we present a general computational method for quantifying parameter uncertainty and assessing parameter identifiability through a parametric bootstrap approach. We demonstrate this approach through examples of compartmental epidemic models with variable complexity, which have been previously employed to study the transmission dynamics and control of various infectious diseases including pandemic influenza, Ebola, and Zika.

2. Methods

2.1. Compartmental Models

Compartmental models are widely used in epidemiological literature as a population-level modeling approach that subdivides the population into classes according to their epidemiological status (1, 20). Compartmental dynamic models are specified by a set of ordinary differential equations and parameters that track the temporal progression of the number of individuals in each of the states of the system (3, 21). Dynamic models follow the general form:

$$\dot{x}_1(t) = f_1(x_1, x_2, \dots, x_h, \Theta)$$
$$\dot{x}_2(t) = f_2(x_1, x_2, \dots, x_h, \Theta)$$

$$\vdots$$

$$\dot{x}_h(t) = f_h(x_1, x_2, \ldots, x_h, \Theta)$$

Where $\dot{x}_i$ is the rate of change of the system states (where i= 1, 2, …, h) and $\Theta = (\theta_1, \theta_2, \ldots, \theta_m)$ is the set of model parameters.

The basic reproductive number (denoted R_0) is often a parameter of interest in epidemiological studies, as it is a measure of potential for a given infectious disease to spread within a population. Mathematically, it is defined as the average number of secondary infections produced by a single index case in a completely susceptible population (22). R_0 represents an epidemic threshold for which values of $R_0 < 1$ indicate a lack of disease spread, and values of $R_0 > 1$ are consistent with epidemic spread. In the midst of an epidemic, R_0 estimates provide insight to the intensity of interventions required to achieve control (23). R_0 is a composite parameter value, as it depends on multiple model parameters (e.g., transmission rate, infectious period), and while R_0 is not directly estimated from the model, it can be calculated by relying on the uncertainty of individual parameters.

A simple and commonly utilized compartmental model is the SEIR (susceptible-exposed-infectious-removed) model (1). We apply our methodology to this low-complexity model and work through increasingly more complex models as we demonstrate the approach for assessing parameter identifiability.

2.1.1. Model 1: Simple SEIR (Pandemic Influenza)

We analyze a simple compartmental transmission model that consists of 4 parameters and 4 states (Figure 1). We apply this model to the context of the 1918 influenza pandemic in San Francisco, California (23). Individuals in the model are classified as susceptible (S), exposed (E), infectious (I), or recovered (R) (1). We assume constant population size, so S + E + I + R = N, where N is the total population size. Susceptible individuals progress to the exposed class at rate $\beta I(t)/N$, where β is the transmission rate, and $I(t)/N$ is the probability of random contact with an infectious individual. Exposed, or latent, individuals move to the infectious class at rate k, where $1/k$ is the average latent period. Infectious individuals recover (move to recovered class) at rate γ, where $1/\gamma$ corresponds to the average infectious period.

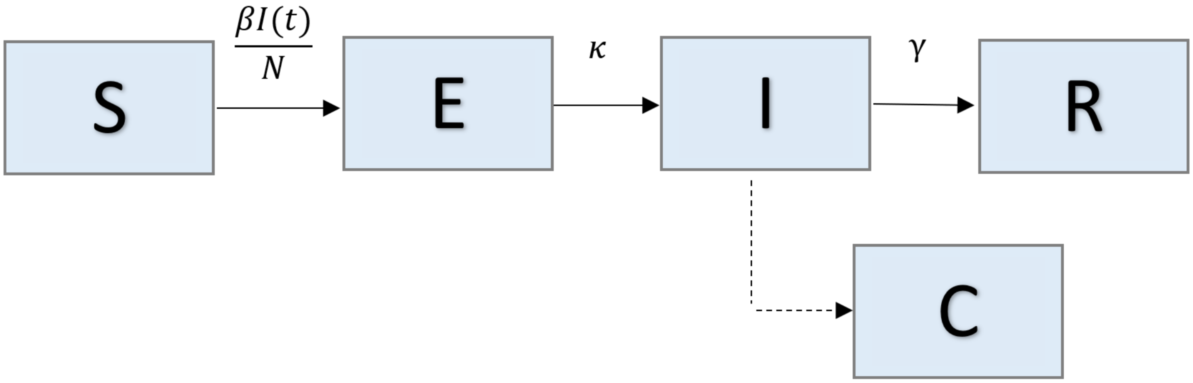


Figure 1. Model 1: Simple SEIR – Population is divided into 4 classes: susceptible (S), exposed (E), infectious (I), and recovered/removed (R). Class C represents the auxiliary variable C(t) and tracks the cumulative number of infectious individuals from the start of the outbreak. This is presented as a dashed line, as it is not a state of the system of equations, but simply a class to track the cumulative incidence cases; meaning, individuals from the population are not moving to class C. Parameter(s) above arrows denote the rate individuals move between classes. Parameter descriptions and values are found in Table 1.

The transmission process can be modeled using the following system of ordinary differential equations (where the dot denotes time derivative):

$$\begin{cases} \dot{S}(t) = -\beta S(t)I(t)/N \\ \dot{E}(t) = \beta S(t)I(t)/N - kE(t) \\ \dot{I}(t) = kE(t) - \gamma I(t) \\ \dot{R}(t) = \gamma I(t) \\ \dot{C}(t) = kE(t) \end{cases}$$

The auxiliary variable C(t) tracks the cumulative number of infectious individuals from the start of the outbreak. It is not a state of the system of equations, but simply a class to track the cumulative incidence cases; meaning, individuals from the population are not moving to class C. The number of new infections, or the incidence curve, is given by $\dot{C}(t)$.

For this model, there is only one class contributing to new infections (I), so R_0, or the basic reproductive number, is simply the product of the transmission rate and the average infectious period: $R_0 = \frac{\beta}{\gamma}$.

2.1.2. Model 2: SEIR with asymptomatic and hospitalized/diagnosed and reported

We use a simplified version of a complex SEIR model that consists of 8 parameters and 6 system states (Figure 2). This model was originally developed for studying the transmission dynamics of the 1918 influenza pandemic in Geneva, Switzerland (24). In the model, individuals are classified as susceptible (S), exposed (E), clinically ill and infectious (I), asymptomatic and partially infectious (A), hospitalized/diagnosed and reported (J), or recovered (R). Hospitalized individuals are assumed to be as infectious as individuals in the I class. Again, constant population size is assumed, so $S + E + I + A + J + R = N$. Susceptible individuals progress to the exposed class at rate $\beta[I(t) + J(t) + qA(t)]/N$, where β is the transmission rate, and q is a reduction factor of transmissibility in the asymptomatic class ($0 < q < 1$). A proportion, ρ, of exposed/latent individuals ($0 < \rho < 1$) become clinically infectious at rate k, while the rest (1- ρ) become partially infectious and asymptomatic at the same rate k. Asymptomatic cases progress to the recovered class at rate γ_1. Clinically ill and infectious individuals are diagnosed at a rate α or recover without being diagnosed at rate γ_1. Diagnosed individuals recover at rate γ_2.

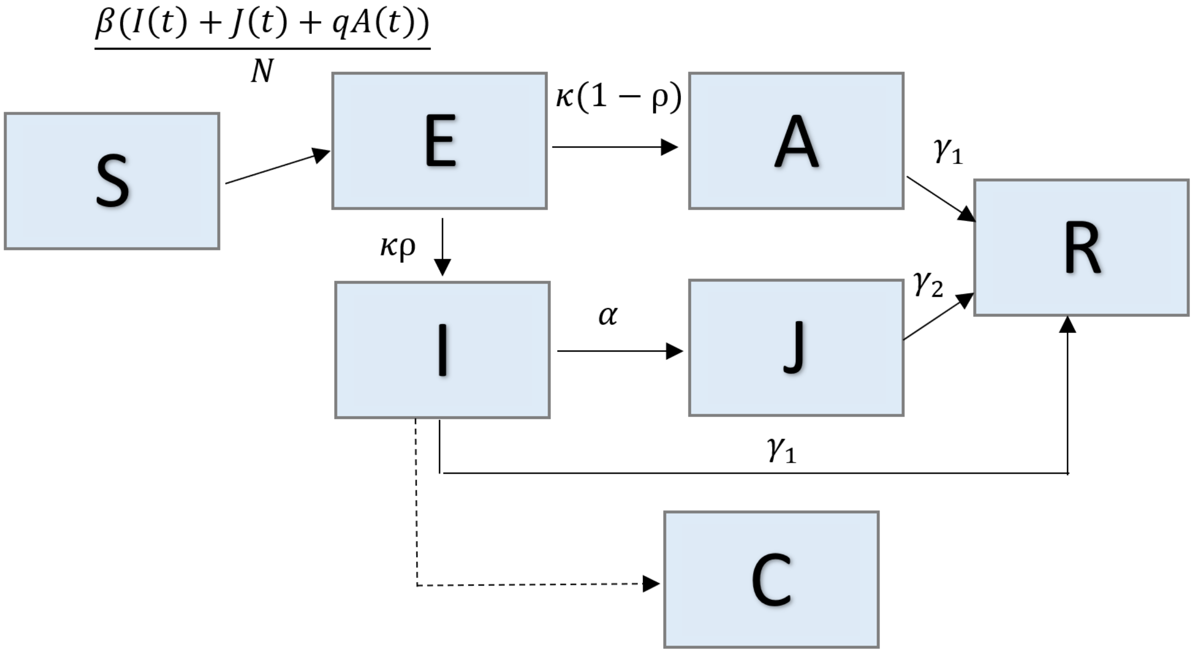


Figure 2. Model 2: SEIR with asymptomatic and hospitalized/diagnosed and reported – Population is divided into 6 classes: susceptible (S), exposed (E), clinically ill and infectious (I), asymptomatic and partially infectious (A), hospitalized/diagnosed and reported (J), and recovered (R). Class C represents the auxiliary variable C(t) and tracks the cumulative number of newly infectious individuals. Parameter(s) above (or to the left of) arrows denote the rate individuals move between classes. Parameter descriptions and values are found in Table 2.

The transmission process can be modeled using the following system of ordinary differential equations:

$$\begin{cases} \dot{S}(t) = -\beta S(t)[I(t)+J(t)+qA(t)]/N \\ \dot{E}(t) = \beta S(t)[I(t)+J(t)+qA(t)]/N - kE(t) \\ \dot{A}(t) = k(1-\rho)E(t) - \gamma_1 A(t) \\ \dot{I}(t) = k\rho E(t) - (\alpha+\gamma_1)I(t) \\ \dot{J}(t) = \alpha I(t) - \gamma_2 J(t) \\ \dot{R}(t) = \gamma_1(A(t)+I(t)) + \gamma_2 J(t) \\ \dot{C}(t) = \alpha I(t) \end{cases}$$

In the above system, C(t) represents the cumulative number of diagnosed/reported cases from the start of the outbreak, and $\dot{C}(t)$ is the incidence curve of diagnosed cases.

For this model, there are three classes contributing to new infections (A, I, J), so the reproductive number is the sum of the contributions from each of these classes: $R_0 = R_0^A + R_0^I + R_0^J$, where:

R_0^A = (fraction of asymptomatic cases) x (transmission rate)
x (relative transmissibility from asymptomatic cases)
x (mean time in asymptomatic class)

R_0^I = (fraction of symptomatic cases) x (transmission rate)
x (mean time in clinically infectious class)

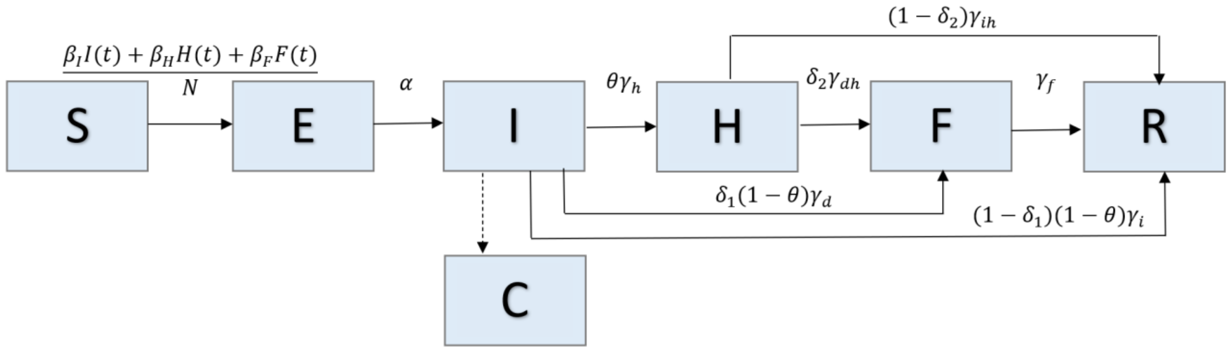
R_0^J = (fraction of symptomatic cases that are hospitalized) x (transmission rate)
x (mean time in hospital) (24)

Here, $R_0 = \beta[(1-\rho)(\frac{q}{\gamma_1}) + \rho(\frac{1}{\gamma_1+\alpha} + \frac{\alpha}{(\gamma_1+\alpha)\gamma_2})]$.

2.1.3. Model 3: The Legrand et al. Model (Ebola)

We analyze an Ebola transmission model (25) comprised of 15 parameters and 6 states (Figure 3). This model subdivides the infectious population into three stages to account for transmission in three settings: community, hospital, and unsafe burial ceremonies. Individuals are classified as susceptible (S), exposed (E), infectious in the community (I), infectious in the hospital (H), infectious after death at funeral (F), or recovered/removed (R). Constant population size is assumed, so S + E + I + H + F + R = N. Susceptible individuals progress to the exposed class at rate $(\beta_I I(t) + \beta_H H(t) + \beta_F F(t))/N$ where β_I, β_H, and β_F represent the transmission rates in the community, hospital, and at funerals, respectively. Exposed individuals become infectious at rate α. A proportion, $0 < \theta < 1$, of infectious individuals are hospitalized at rate γ_h. Of the proportion of infectious individuals that are not hospitalized (1-θ), a proportion, $0 < \delta_1 < 1$, move to the funeral class at rate γ_d, and the rest (1- δ_1) move to the recovered/removed class at rate γ_i. A proportion, $0 < \delta_2 < 1$, of hospitalized individuals progress to funeral class at rate $\gamma_{dh} = \frac{1}{\frac{1}{\gamma_d} - \frac{1}{\gamma_h}}$.

The remaining proportion (1- δ$_2$) are recovered/removed at rate $\gamma_{ih} = \frac{1}{\frac{1}{\gamma_i} - \frac{1}{\gamma_h}}$. δ$_1$ and δ$_2$ are calculated such that δ represents the case fatality ratio (Table 3). Individuals in the funeral class are removed at rate γ$_f$.



$\frac{\beta_I I(t) + \beta_H H(t) + \beta_F F(t)}{N}$
α
$\theta\gamma_h$
$(1-\delta_2)\gamma_{ih}$
$\delta_2\gamma_{dh}$
γ_f
S
E
I
H
F
R
$\delta_1(1-\theta)\gamma_d$
$(1-\delta_1)(1-\theta)\gamma_i$
C

Figure 3. Model 3: The Legrand *et al.* Model – Population is divided into 6 classes: susceptible (S), exposed (E), infectious in the community (I), infectious in the hospital (H), infectious after death at funeral (F), or recovered/removed (R). Class C represents the auxiliary variable C(t) and tracks the cumulative number of newly infectious individuals. Parameter(s) above arrows denote the rate that individuals move between classes. Parameter descriptions and values are found in Table 3.

The transmission process is modeled by the following set of ordinary differential equations:

$$\begin{cases} \dot{S}(t) = -S(t)[\beta_I I(t) + \beta_H H(t) + \beta_F F(t)]/N \\ \dot{E}(t) = S(t)[\beta_I I(t) + \beta_H H(t) + \beta_F F(t)]/N - \alpha E(t) \\ \dot{I}(t) = \alpha E(t) - [\theta\gamma_h + \delta_1(1-\theta)\gamma_d + (1-\delta_1)(1-\theta)\gamma_i]I(t) \\ \dot{H}(t) = \theta\gamma_h I(t) - [(1-\delta_2)\gamma_{ih} + \delta_2\gamma_{dh}]H(t) \\ \dot{F}(t) = \delta_1(1-\theta)\gamma_d I(t) + \delta_2\gamma_{dh}H(t) - \gamma_f F(t) \\ \dot{R}(t) = (1-\delta_1)(1-\theta)\gamma_i I(t) + (1-\delta_2)\gamma_{ih}H(t) + \gamma_f F(t) \\ \dot{C}(t) = \alpha E(t) \end{cases}$$

Here, C(t) represents the cumulative number of all infectious individuals, and $\dot{C}(t)$ is the incidence curve for infectious cases.

The basic reproductive number is the sum of the contributions from each of the infectious classes (I, H, F): $R_0 = R_0^I + R_0^H + R_0^F$, where:

R_0^I = (transmission rate in the community) x (mean time in infectious class)

R_0^H = (fraction of hospitalized cases) x (transmission rate in the hospital)
x (mean time in hospital class)

R_0^F = (fraction of cases that have traditional burial ceremonies) x (transmission rate at funerals)
x (mean time in funeral class)

Here, $R_0 = \frac{\beta_I}{\Delta} + \frac{\frac{\gamma_h \theta}{\gamma_{dh}\delta_2 + \gamma_{ih}(1-\delta_2)}\beta_H}{\Delta} + \frac{\gamma_d \delta_1 (1-\theta)\beta_F}{\gamma_f \Delta} + \frac{\gamma_{dh}\gamma_h \delta_2 \theta \beta_F}{\gamma_f(\gamma_{ih}(1-\delta_2) + \gamma_{dh}\delta_2)\Delta}$,
where $\Delta = \gamma_h \theta + \gamma_d (1-\theta)\delta_1 + \gamma_i (1-\theta)(1-\delta_1)$ (25).

2.1.4. Model 4: Zika Model with human and mosquito populations

The last example is a compartmental model of Zika transmission dynamics that includes 16 parameters and 9 states and incorporates transmission between two populations – humans and vectors (Figure 4). This model was designed to investigate the impact of both mosquito-borne and sexually transmitted (human-to-human) routes of infection for cases of Zika virus (26). In the human population, individuals are classified as susceptible (S_h), asymptomatically infected (A_h), exposed (E_h), symptomatically infectious (I_{h1}), convalescent (I_{h2}), or recovered (R_h). The mosquito, or vector, population is broken into susceptible (S_v), exposed (E_v), and infectious (I_v) classes. Note that the subscript 'h' is used for humans and 'v' is used for vectors. Constant population size is assumed in both populations, so $S_h + A_h + E_h + I_{h1} + I_{h2} + R_h = N_h$ and $S_v + E_v + I_v = N_v$.

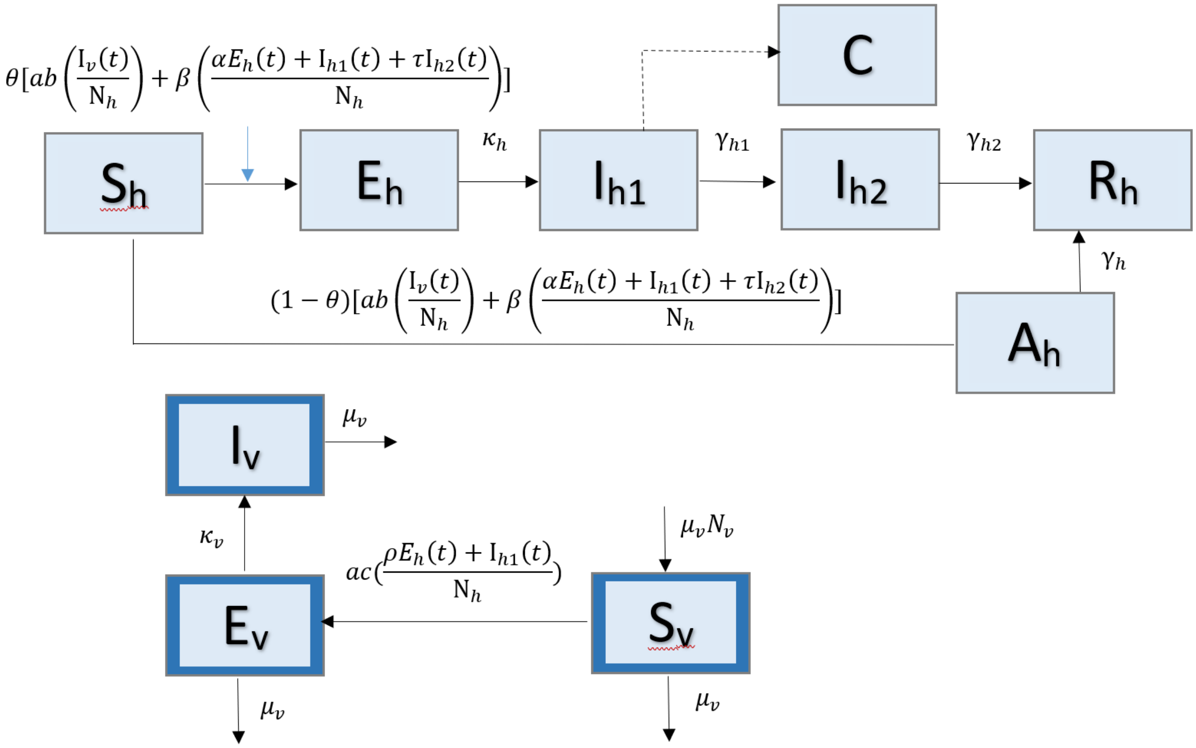

$\theta[ab\left(\frac{\mathrm{I}_v(t)}{\mathrm{N}_h}\right) + \beta\left(\frac{\alpha E_h(t) + \mathrm{I}_{h1}(t) + \tau\mathrm{I}_{h2}(t)}{\mathrm{N}_h}\right)]$
C
S_h
E_h
κ_h
I_{h1}
γ_{h1}
I_{h2}
γ_{h2}
R_h
γ_h
$(1-\theta)[ab\left(\frac{\mathrm{I}_v(t)}{\mathrm{N}_h}\right) + \beta\left(\frac{\alpha E_h(t) + \mathrm{I}_{h1}(t) + \tau\mathrm{I}_{h2}(t)}{\mathrm{N}_h}\right)]$
A_h
I_v
μ_v
κ_v
$\mu_v N_v$
$ac(\frac{\rho E_h(t) + \mathrm{I}_{h1}(t)}{\mathrm{N}_h})$
E_v
S_v
μ_v
μ_v

Figure 4. Model 4: Zika Model with human and mosquito populations – The human population (subscript h) is divided into 5 classes: susceptible (S_h), asymptomatically infected (A_h), exposed (E_h), symptomatically infectious (I_{h1}), convalescent (I_{h2}), or recovered (R_h). Class C represents the auxiliary variable C(t) and tracks the cumulative number of newly infectious individuals. The mosquito, or vector, population (subscript v; outlined in dark blue) is divided into 3 classes: susceptible (S_v), exposed (E_v), and infectious (I_v) classes. Parameter(s) above arrows denote the rate individuals/vectors move between classes. Parameter descriptions and values are found in Table 4.

A proportion $0 < \theta < 1$ of susceptible humans move to the exposed class at rate $ab(\mathrm{I}_v(t)/\mathrm{N}_h) + \beta[(\alpha E_h(t) + \mathrm{I}_{h1}(t) + \tau\mathrm{I}_{h2}(t))/\mathrm{N}_h)]$ where a is the mosquito biting rate, b is the transmission probability from an infectious mosquito to a susceptible human, β is the transmission rate between humans, α is the relative (human-to-human) transmissibility from exposed humans to susceptible, and τ is the relative transmissibility from convalescent humans compared to susceptible. Exposed individuals progress to symptomatically infectious at rate κ_h and then progress to the convalescent stage at rate γ_{h1}. Convalescent individuals recover at rate γ_{h2}. The remaining proportion of susceptible individuals (1 - θ) become asymptomatically infected at the same rate, $ab(\mathrm{I}_v(t)/\mathrm{N}_h) + \beta[(\alpha E_h(t) + \mathrm{I}_{h1}(t) + \tau\mathrm{I}_{h2}(t))/\mathrm{N}_h]$. Asymptomatic humans recover at rate γ_h and do not contribute to new infections in this model.

Susceptible mosquitos move to the exposed class at rate $ac[(\rho E_h(t) + \mathrm{I}_{h1}(t))/\mathrm{N}_h]$, where c is the transmission probability from a symptomatically infectious human to a susceptible mosquito, and ρ is the relative human-to-mosquito transmission probability from exposed humans to symptomatically infected. Exposed mosquitos become infectious at rate κ_v. Mosquitos also leave the population at rate μ_v, where $1/\mu_v$ is the mosquito lifespan.

The transmission process, including both populations, is represented by the set of differential equations below:

$$\begin{cases}
\dot{S}_h(t) = -ab(\mathrm{I}_v(t)/\mathrm{N}_h)\mathrm{S}_h(t) - \beta[(\alpha E_h(t) + \mathrm{I}_{h1}(t) + \tau\mathrm{I}_{h2}(t))/\mathrm{N}_h]\mathrm{S}_h(t) \\
\dot{E}_h(t) = \theta[ab(\mathrm{I}_v(t)/\mathrm{N}_h)\mathrm{S}_h(t) + \beta[(\alpha E_h(t) + \mathrm{I}_{h1}(t) + \tau\mathrm{I}_{h2}(t))/\mathrm{N}_h]\mathrm{S}_h(t)] - \kappa_h E_h(t) \\
\dot{I}_{h1}(t) = \kappa_h E_h(t) - \gamma_{h1}\mathrm{I}_{h1}(t) \\
\dot{I}_{h2}(t) = \gamma_{h1}\mathrm{I}_{h1}(t) - \gamma_{h2}\mathrm{I}_{h2}(t) \\
\dot{A}_h(t) = (1-\theta)[ab(\mathrm{I}_v(t)/\mathrm{N}_h)\mathrm{S}_h(t) + \beta[(\alpha E_h(t) + \mathrm{I}_{h1}(t) + \tau\mathrm{I}_{h2}(t))/\mathrm{N}_h]\mathrm{S}_h(t)] - \gamma_h A_h(t) \\
\dot{R}_h(t) = \gamma_{h2}\mathrm{I}_{h2}(t) + \gamma_h A_h(t) \\
\dot{S}_v(t) = \mu_v N_v - ac[(\rho E_h(t) + \mathrm{I}_{h1}(t))/\mathrm{N}_h] * \mathrm{S}_v(t) - \mu_v\mathrm{S}_v(t) \\
\dot{E}_v(t) = ac[(\rho E_h(t) + \mathrm{I}_{h1}(t))/\mathrm{N}_h] * \mathrm{S}_v(t) - (\kappa_v + \mu_v)\mathrm{E}_v(t) \\
\dot{I}_v(t) = \kappa_v\mathrm{E}_v(t) - \mu_v\mathrm{I}_v(t) \\
\dot{C}(t) = \kappa_h E_h(t)
\end{cases}$$

C(t) represents the cumulative number of symptomatically infectious human cases, and $\dot{C}(t)$ contains the incidence curve for symptomatic human cases.

For this example, we have two transmission processes to consider when calculating R_0: sexual transmission (R_{hh}) and mosquito-borne (R_{hv}). The human population has three classes contributing to new infections: exposed, symptomatically infectious, and convalescent, so:

$$R_{hh} = \frac{\alpha\theta\beta}{\kappa_h} + \frac{\theta\beta}{\gamma_{h1}} + \frac{\tau\theta\beta}{\gamma_{h2}}$$

The mosquito population only has one infectious class (I_v); the reproductive number is given by:

$$R_{hv} = \sqrt{[\frac{a^2 b \rho c m \theta}{\kappa_h \mu_v} + \frac{a^2 b c m \theta}{\gamma_{h1} \mu_v}] * \frac{\kappa_v}{\kappa_v + \mu_v}} \, .$$

The overall basic reproductive number, considering both transmission routes, is given by the following equation (26):

$$R_0 = \frac{R_{hh} + \sqrt{R_{hh}^2 + 4R_{hv}^2}}{2}$$

2.2 Simulated data

For each model we simulate 200 epidemic datasets (directly from the corresponding set of ordinary differential equations) with Poisson error structure using the daily time series data of case incidence, or total number of new cases daily. Parameters for each model are set at values based on their corresponding application: the 1918 influenza pandemic in San Francisco (Model 1) (23), 1918 pandemic influenza in Geneva (Model 2) (24), 1995 Ebola in Congo (Model 3) (25), and 2016 Zika in the Americas (Model 4) (26). As explained below, the simulated data are generated using a bootstrap approach, and we then use these data to study parameter identifiability within a realistic parameter space for each model. Parameter descriptions and their corresponding values for each model are given in Tables 1-4.

Table 1. Parameter descriptions and values for Model 1

Parameters	**Description**	**Value**
N	Population size	500000
β	Transmission rate (per day)	0.56
1/κ	Mean latent period (days)	1.9
1/γ	Mean infectious period (days)	4.1
R_0	Basic reproductive number	2.3

Parameter values are consistent with pandemic influenza in San Francisco, 1918 (23).

Table 2. Parameter descriptions and values for Model 2

Parameters	Description	Value
N	Population size	500000
β	Transmission rate (per day)	0.8
1/κ	Latent period (days)	1.9
γ_1	Recovery rate for asymptomatic individuals (1/days)	1/4.1
γ_2	Recovery rate for infectious individuals recovering without hospitalization (1/days)	1/2.3
α	Rate of diagnosis for hospitalized individuals (days)	0.555
ρ	Proportion of latent individuals progressing to infectious class (vs. asymptomatic class)	0.6
q	Reduction factor in transmissibility for asymptomatic cases	0.4
R_0	Basic reproductive number	1.89

Parameter values are consistent with pandemic influenza in Geneva, 1918 (24).

Table 3. Parameter descriptions and values for Model 3

Parameters	Description	Value
N	Population size	200000
β_I	Transmission rate in the community (per day)	0.084
β_H	Transmission rate in the hospital (per day)	0.1134
β_F	Transmission rate at traditional funerals (per day)	1.093
1/α	Incubation period (days)	7
θ	Proportion of cases hospitalized	0.80
$1/\gamma_h$	Time from symptom onset to hospitalization (days)	5
$1/\gamma_d$	Time from symptom onset to death (days)	9.6
$1/\gamma_i$	Time from symptom onset to the end of infectiousness for survivors (days)	10
δ	Case fatality ratio	0.81

δ_1	$\delta_1 = \dfrac{\delta\gamma_i}{\delta\gamma_i + (1-\delta)\gamma_d}$	0.80
δ_2	$\delta_2 = \dfrac{\delta\gamma_{ih}}{\delta\gamma_{ih} + (1-\delta)\gamma_{dh}}$	0.80
$1/\gamma_{ih}$	Infectious period for survivors (days)	5
$1/\gamma_{dh}$	Time from hospitalization to death (days)	4.6
$1/\gamma_f$	Time from death to funeral (days)	2
R_0	Basic reproductive number	2.685

Parameter values are consistent with the 1995 Ebola outbreak in the Democratic Republic of Congo (25).

Table 4. Parameter descriptions and values for Model 4

Parameters	**Description**	**Value**
N_h	Population size (humans)	200000
N_v	Population size (mosquitos)	1000000
a	Mosquito biting rate (number of bites per mosquito per day)	0.5
b	Probability of infection from an infectious mosquito to a susceptible human (per bite)	0.4
β	Transmission rate from symptomatically infected humans to susceptible humans (per day)	0.05
α	Relative human-to-human transmissibility of exposed humans to symptomatic humans	0.6
τ	Relative human-to-human transmissibility of convalescent to symptomatic humans	0.3
ϴ	Proportion of symptomatic infections	0.18
$1/\kappa_h$	Intrinsic incubation period in humans (days)	5
$1/\gamma_{h1}$	Duration of acute phase (days)	5
$1/\gamma_{h2}$	Duration of convalescent phase (days)	20
$1/\gamma_h$	Duration of asymptomatic infection (days)	
$1/\mu_v$	Mosquito lifespan (days)	14

c	Transmission probability from a symptomatically infected human to a susceptible mosquito per bite	0.5
ρ	Relative human-to-mosquito transmission probability of exposed humans to symptomatically infected humans	0.1
$1/\kappa_v$	Extrinsic incubation period in mosquitos (days)	10
R_0	Basic reproductive number	1.486

Parameter values are consistent with the 2016 Zika outbreak in Brazil, Colombia, and El Salvador (26).

2.3. Parameter Estimation

To estimate parameter values, we fit the model to each simulated dataset using nonlinear least squares estimation. The *lsqcurvefit* function in Matlab (Mathworks, Inc.) is used to find the least squares best fit to the data. This process searches for the set of parameters $\hat{\Theta}$= ($\hat{\theta}_1$, $\hat{\theta}_2$,…, $\hat{\theta}_m$) that minimizes the sum of squared differences between the simulated data and the model solution (3). The model solution $f(t_i, \hat{\Theta})$ represents the best fit to the time series data.

For this method, the initial parameter predictions affect the solution for the model as local minima occur. While we know the true parameter values (used to generate the data), this is unrealistic for a real-world modeling scenario. We vary the initial guesses of the parameter values to vary according to a uniform distribution in the range of +/- 0.1 around the true value. Another approach would consist of repeating the least squares fitting procedure several times with different initial parameter guesses and selecting the best model fit.

For each model, the sets of parameters are denoted by Θ_i , where i represents the number of parameters being jointly estimated. We begin with estimating one model parameter, while fixing the rest, and then increase the number of parameters jointly estimated by one until all parameters of interest are included. Population size, N, is always fixed to the true value. Also, while R_0 is not being directly estimated from the model, it is a composite parameter that can be calculated using individual parameter estimates.

For each model described above, we explore parameter identifiability for the following sets of parameters. Here, the symbol ^ is used to indicate an estimated parameter, while the absence of this symbol indicates that the parameter is set to its true value from the simulated data.

(i) Model 1: Simple SEIR

Θ_i: $\Theta_1 = \{\, \hat{\beta}, \kappa, \gamma \,\}$

$\Theta_2 = \{\, \hat{\beta}, \kappa, \hat{\gamma} \,\}$

$\Theta_3 = \{\, \hat{\beta}, \hat{\kappa}, \hat{\gamma} \,\}$

(ii) Model 2: SEIR with asymptomatic and hospitalized/diagnosed and reported

Θ_i: $\Theta_1 = \{\, \hat{\beta}, \kappa, \gamma_1, \gamma_2, \alpha, \rho, q \,\}$

$\Theta_2 = \{\, \hat{\beta}, \kappa, \widehat{\gamma_1}, \gamma_2, \alpha, \rho, q \,\}$

$\Theta_3 = \{\, \hat{\beta}, \kappa, \widehat{\gamma_1}, \gamma_2, \hat{\alpha}, \rho, q \,\}$

$\Theta_4 = \{\, \hat{\beta}, \kappa, \widehat{\gamma_1}, \gamma_2, \hat{\alpha}, \hat{\rho}, q \,\}$

$\Theta_5 = \{\, \hat{\beta}, \kappa, \widehat{\gamma_1}, \gamma_2, \hat{\alpha}, \hat{\rho}, \hat{q} \,\}$

(iii) Model 3: The Legrand Model (Ebola)

Θ_i: $\Theta_1 = \{\, \hat{\beta}_I, \beta_H, \beta_F, \alpha, \theta, \gamma_h, \gamma_d, \gamma_i, \delta, \gamma_{ih}, \gamma_{dh}, \gamma_f \,\}$

$\Theta_2 = \{\, \hat{\beta}_I, \hat{\beta}_H, \beta_F, \alpha, \theta, \gamma_h, \gamma_d, \gamma_i, \delta, \gamma_{ih}, \gamma_{dh}, \gamma_f \,\}$

$\Theta_3 = \{\, \hat{\beta}_I, \hat{\beta}_H, \hat{\beta}_F, \alpha, \theta, \gamma_h, \gamma_d, \gamma_i, \delta, \gamma_{ih}, \gamma_{dh}, \gamma_f \,\}$

$\Theta_4 = \{\, \hat{\beta}_I, \hat{\beta}_H, \hat{\beta}_F, \alpha, \theta, \hat{\gamma}_h, \gamma_d, \gamma_i, \delta, \gamma_{ih}, \gamma_{dh}, \gamma_f \,\}$

$\Theta_5 = \{\, \hat{\beta}_I, \hat{\beta}_H, \hat{\beta}_F, \alpha, \theta, \hat{\gamma}_h, \hat{\gamma}_d, \gamma_i, \delta, \gamma_{ih}, \gamma_{dh}, \gamma_f \,\}$

$\Theta_6 = \{\, \hat{\beta}_I, \hat{\beta}_H, \hat{\beta}_F, \alpha, \theta, \hat{\gamma}_h, \hat{\gamma}_d, \hat{\gamma}_i, \delta, \gamma_{ih}, \gamma_{dh}, \gamma_f \,\}$

$\Theta_7 = \{\, \hat{\beta}_I, \hat{\beta}_H, \hat{\beta}_F, \alpha, \theta, \hat{\gamma}_h, \hat{\gamma}_d, \hat{\gamma}_i, \delta, \gamma_{ih}, \gamma_{dh}, \hat{\gamma}_f \,\}$

(iv) Model 4: Zika model with human and mosquito populations

Θ_i: $\Theta_1 = \{\, a, b, \hat{\beta}, \alpha, \tau, \theta, \kappa_h, \gamma_{h1}, \gamma_{h2}, \gamma_h, \mu_v, c, \rho, \kappa_v \,\}$

$\Theta_2 = \{\, a, b, \hat{\beta}, \alpha, \tau, \theta, \kappa_h, \hat{\gamma}_{h1}, \gamma_{h2}, \gamma_h, \mu_v, c, \rho, \kappa_v \,\}$

$\Theta_3 = \{\, a, b, \hat{\beta}, \alpha, \tau, \theta, \kappa_h, \hat{\gamma}_{h1}, \hat{\gamma}_{h2}, \gamma_h, \mu_v, c, \rho, \kappa_v \,\}$

$$\Theta_4 = \{ a, b, \hat{\beta}, \alpha, \tau, \theta, \kappa_h, \hat{\gamma}_{h1}, \hat{\gamma}_{h2}, \hat{\gamma}_h, \mu_v, c, \rho, \kappa_v \}$$

$$\Theta_5 = \{ a, b, \hat{\beta}, \hat{\alpha}, \tau, \theta, \kappa_h, \hat{\gamma}_{h1}, \hat{\gamma}_{h2}, \hat{\gamma}_h, \mu_v, c, \rho, \kappa_v \}$$

$$\Theta_6 = \{ a, b, \hat{\beta}, \hat{\alpha}, \hat{\tau}, \theta, \kappa_h, \hat{\gamma}_{h1}, \hat{\gamma}_{h2}, \hat{\gamma}_h, \mu_v, c, \rho, \kappa_v \}$$

2.4. Bootstrapping Method

We use the parametric bootstrap approach (3, 27, 28) for simulating the error structure around the deterministic model solution in order to evaluate parameter identifiability. This computational approach involves repeatedly sampling observations from the best-fit model solution. Here we use a Poisson error structure, which is the most popular distribution for modeling count data [3]. The step-by-step approach to quantify parameter uncertainty is as follows:

1. Obtain the deterministic model solution (total daily incidence series) using nonlinear least-squares estimation (*Section 2.3*).
2. Generate S replicate datasets, assuming Poisson error structure:
   Using the deterministic model solution $f(t_i, \hat{\Theta})$, generate S (for our examples, S=200) replicate simulated datasets $f_S^*(t_i, \hat{\Theta})$. To incorporate Poisson error structure, we use the incidence curve, $\dot{C}(t)$, as follows. For each time point t, we generate a new incidence value using a Poisson random variable with mean=$\dot{C}(t)$. This new set of data represents an incidence curve for the system, assuming the time series follows a Poisson distribution centered on the mean at time points t_i.
3. Re-estimate model parameters: For each simulated dataset, derive the best-fit estimates for the parameter set using least-squares fitting (*Section 2.3*). This results in S estimated parameter sets: $\hat{\Theta}_i$ where i=1, 2, …, S.
4. Characterize empirical distributions and construct confidence intervals: Using the set of S parameter estimates, we can characterize the empirical distribution and construct confidence intervals for each estimated parameter. Also, for each set of estimated parameters, R_0 is calculated to obtain a distribution of R_0 values as well.

2.5. Parameter Identifiability

When a model parameter is identifiable from available data, its confidence interval lies in a finite range of values (29, 30). Using the bootstrapping method outlined in *Section 2.4*, we obtain 95% confidence intervals from the distributions of each estimated parameter. A small confidence interval with a finite range of values indicates that the parameter can be precisely identified, while a wider range could be indicative of lack of identifiability. To assess the level of bias of the estimates, we calculate the mean squared error (MSE) for each parameter. MSE is calculated as: $MSE = \frac{1}{S}\sum_{i=1}^{S}(\theta - \widehat{\theta}_i)^2$ where θ represents the true parameter value (in the simulated data), and $\widehat{\theta}_i$ represents the estimated value of the parameter for the ith bootstrap realization.

When a parameter can be estimated with low MSE and narrow confidence, this suggests that the parameter is identifiable from the model. On the other hand, larger confidence intervals or larger MSE values may be suggestive of non-identifiability.

3. Results

3.1. Model 1: Simple SEIR

Supplemental Figures S1-S3 illustrate the empirical distributions of the estimated parameters, where Figure S1 represents the results for $\widehat{\Theta}_1$ (β only), Figure S2 for $\widehat{\Theta}_2$ (β and γ), and Figure S3 for $\widehat{\Theta}_3$ (β, γ, and κ). The figures also show the original simulated data and the 200 simulated datasets for each estimated parameter set.

Estimating only β (Θ_1), results in precise (small confidence interval range) and unbiased (small MSE) estimates of β. Similarly, estimating β and γ (Θ_2) provides precise and unbiased estimates for both parameters. The precision of the estimates can be seen in Figure 5: the confidence intervals for the estimates (represented by red vertical lines) remain close to the true parameter value (blue horizontal dotted line). The MSE plot (Figure 6) shows an MSE value of $< 10^{-7}$ for β in Θ_1 and values of $< 10^{-4}$ for both β and γ in Θ_2.

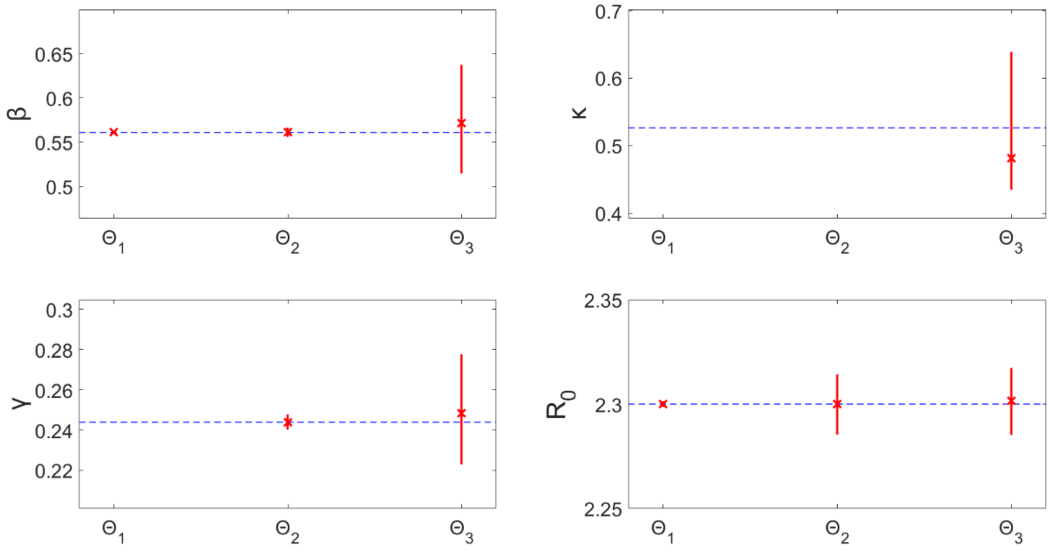


Figure 5. Model 1 – 95% confidence intervals (vertical red lines) for the distributions of each estimated parameter obtained from the 200 realizations of the simulated datasets. Mean estimated parameter value is denoted by a red x, and the true parameter value is represented by the blue dashed horizontal line. Θ_i denotes the estimated parameter set, where i indicates the number of parameters being jointly estimated.

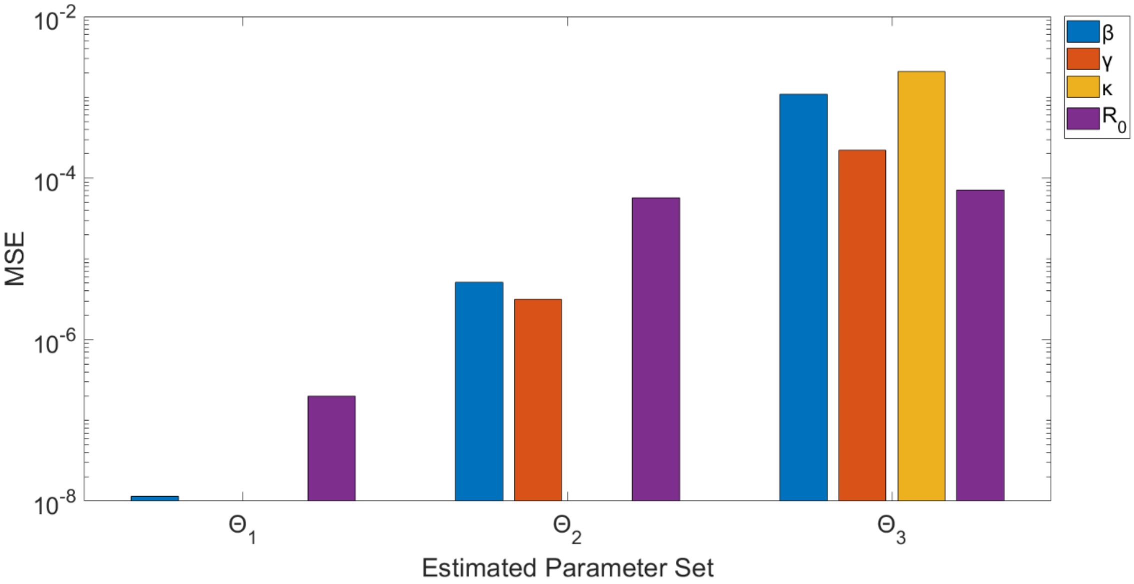


Figure 6. Model 1 – Mean squared error (MSE) of the distribution of parameter estimates (200 realizations) for each estimated parameter set Θ_i, where i indicates the number of parameters being jointly estimated. Note that the y-axis (MSE) is represented with a logarithmic scale.

Simultaneously estimating all 3 parameters, β, κ, and γ (Θ_3), results in wider confidence intervals and larger MSE than the two previous subsets. The confidence intervals for β (0.516, 0.636) and γ (0.223, 0.277) have a narrow range and enclose the true values of the parameters. The MSE for these two are larger compared to the previous subsets, though all MSE values are $< 10^{-2}$. The confidence interval for κ has a slightly larger range (0.440, 0.613), though this correlates with a small latent period difference of less than a day. Also, the MSE for κ is comparable to the other parameters. This indicates that all three parameters can be identified from daily incidence data of the epidemic curve with Poisson error structure.

Moreover, R_0 can be estimated precisely with unbiased results. Despite the larger confidence intervals for the other parameters estimated in Θ_3 (compared to Θ_1, Θ_2), the range around R_0 is still very precise: (2.286, 2.317). Similarly, MSE for R_0 is $< 10^{-4}$ for all runs. This indicates that the estimates of R_0 are robust to variation or bias in the other parameter estimates – we will continue to explore this theme in the proceeding models.

3.2. Model 2: SEIR with asymptomatic and hospitalized/diagnosed and reported

Estimating β only (Θ_1) or β and γ_1 (Θ_2) provides precise estimates with small MSE (Figures 7 & 8). For each Θ_i (where i > 2), each additional parameter being estimated corresponds with, on average, a larger confidence interval range and higher MSE for each estimated parameter. Essentially, for each parameter, the uncertainty grows with the number of other parameters being jointly estimated. Θ_3, estimating β, γ_1, and α, provides estimates of β and γ_1 with relatively small confidence ranges (95% CI: (0.717, 0.851), (0.192, 0.286), respectively) and MSE values (MSE= 0.0016, $7.15*10^{-4}$, respectively); however, estimates for α produce a wider range of values (0.386, 0.748), as well as an MSE value over 5 times higher than the other parameters (MSE=0.0089), though still < 10^{-2}.

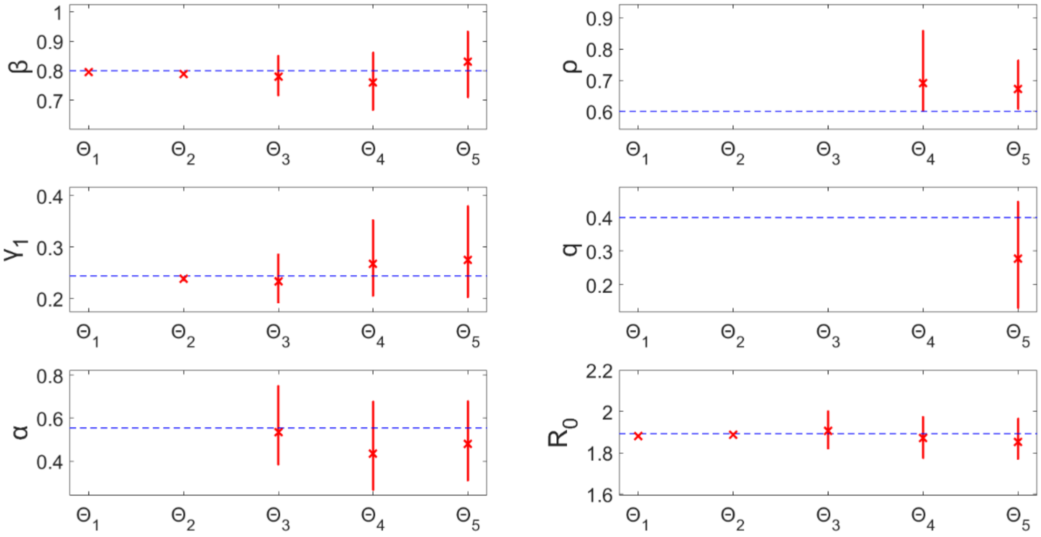


Figure 7. Model 2 – 95% confidence intervals (vertical red lines) for the parameter estimate distributions obtained from the 200 realizations of the simulated datasets. Mean estimated parameter value is denoted by red x, and the true parameter value is represented by the blue dashed horizontal line. Θ_i denotes the estimated parameter set, where i indicates the number of parameters being jointly estimated.

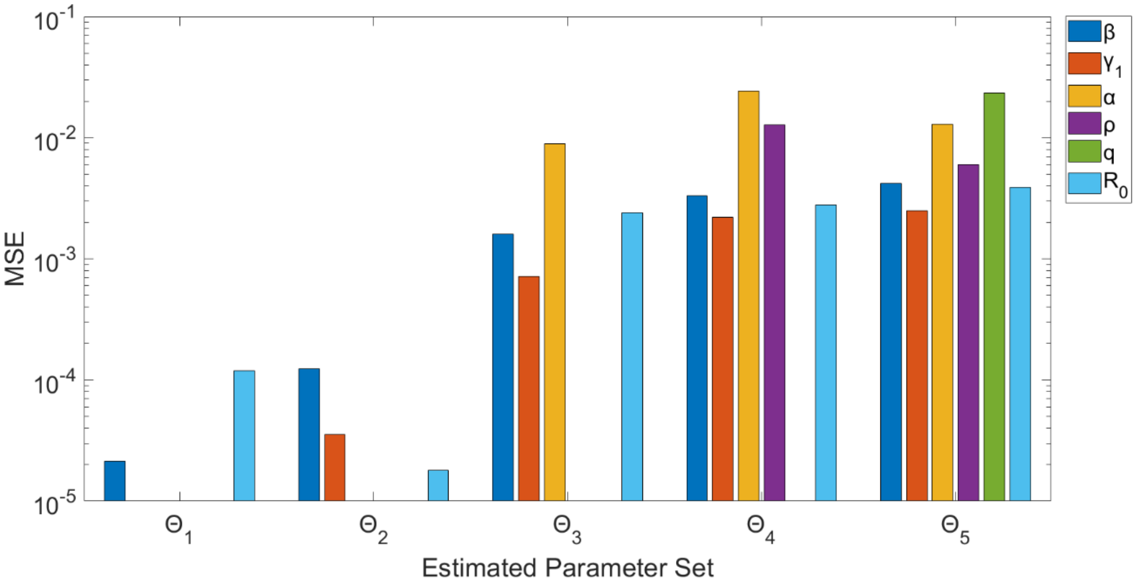


Figure 8. Model 2 – Mean squared error (MSE) of the distribution of parameter estimates (200 realizations) for each estimated parameter set Θ_i, where i indicates the number of parameters being jointly estimated. Note that the y-axis (MSE) is represented with a logarithmic scale.

Results for Θ_4 and Θ_5 indicate that none of the parameters can be well-identified from case incidence data while simultaneously estimating > 3 parameters. For each, multiple parameters have MSE values > 10^{-2} (Figure 8), and the confidence intervals are comparatively wide. Additionally, the confidence intervals for ρ (Θ_4: (0.602, 0.858); Θ_5: (0.608, 0.763)) do not include the true value of 0.60.

Looking at confidence intervals and MSE (Figures 7 & 8) for R_0, we find again that R_0 is identifiable across each Θ_i. The confidence intervals for R_0 all have a range < 0.2, and the MSE values for each Θ_i are < 10^{-2}. These R_0 results are consistent with those in Model 1, despite the identifiability issues of other parameters seen here in Model 2. This is an important result, indicating that even when identifiability issues exist in other model parameters, we can still provide reliable estimates of R_0 without having to know the true values of the other parameters. It also shows that while noise in the data may affect parameter estimation for some parameters, composite parameters, like R_0, can still be accurately calculated from the same data.

3.3. Model 3: The Legrand Model (Ebola)

Estimated parameter sets Θ_1 and Θ_2 (β_I only, β_I and β_H respectively) result in unbiased (MSE < 10^{-3}), precise estimates of the parameters (Figures 9 & 10). However, when jointly estimating all three β values (Θ_3), only β_I is identifiable – the confidence interval is a finite range: (0.038, 0.102) and the estimates are unbiased (MSE= 2.71*10^{-4}). Parameters β_H (0, 0.614) and β_F (0.097, 1.341) both have wide confidence intervals indicating uncertainty suggestive of non-identifiability. Estimating four parameters (Θ_4), only β_H is identifiable with a small range and bias; whereas, the remaining three parameter estimates have larger confidence intervals (Figure 9).

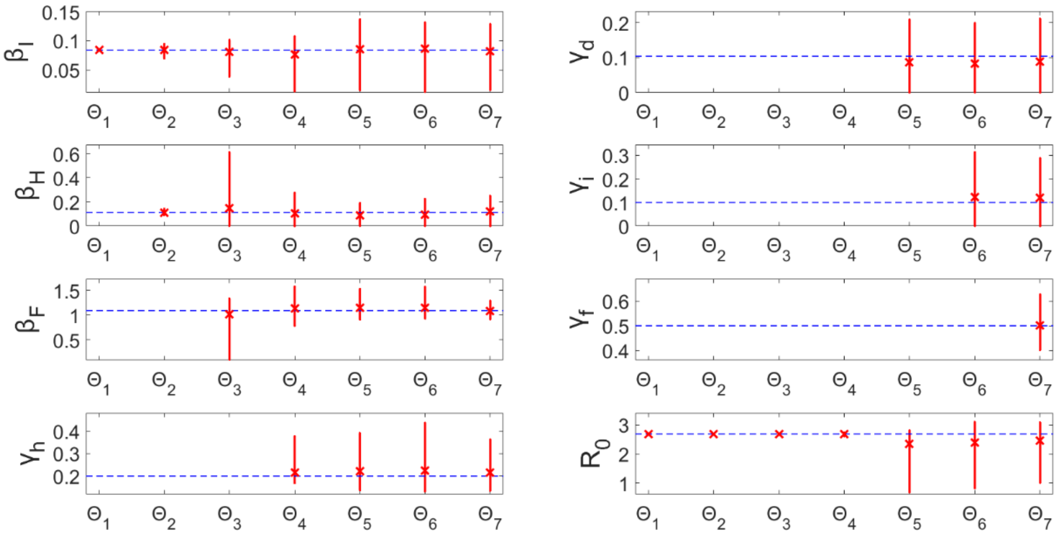


Figure 9. Model 3 – 95% confidence intervals (vertical red lines) for the parameter estimate distributions obtained from the 200 realizations of the simulated datasets. Mean estimated parameter value is denoted by red x, and the true parameter value is represented by the blue horizontal line. Θ_i denotes the estimated parameter set, where i indicates the number of parameters being jointly estimated.

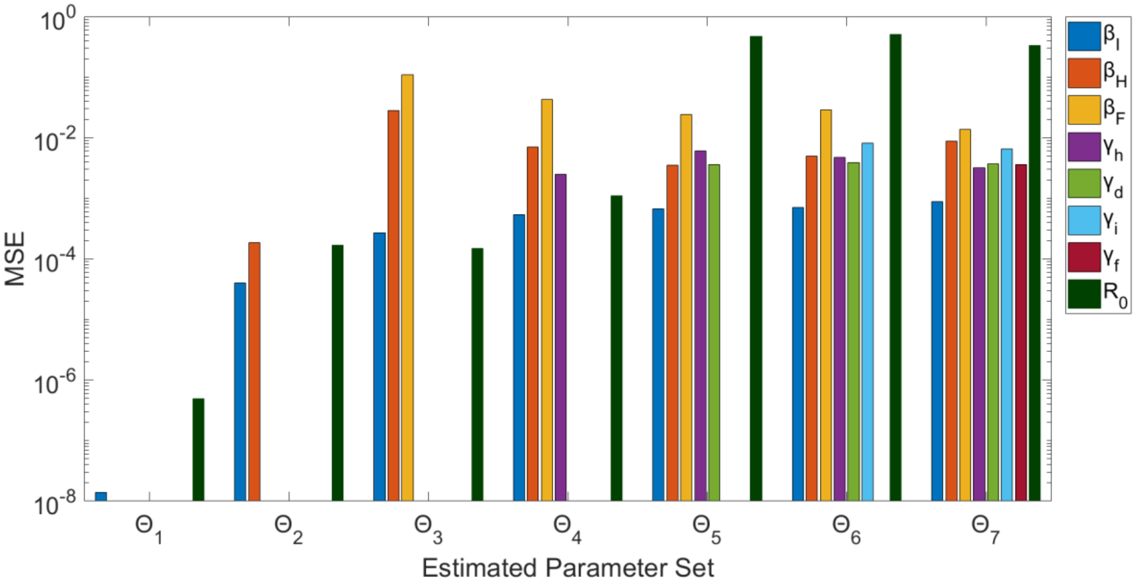


Figure 10. Model 3 – Mean squared error (MSE) of the distribution of parameter estimates (200 realizations) for each estimated parameter set Θ_i, where i indicates the number of parameters being jointly estimated. Note that the y-axis (MSE) is represented with a logarithmic scale.

For Θ_i where i > 4, none of the parameters can be identified from the model/data. Each parameter (for runs $\Theta_5 – \Theta_7$) has either a large confidence range and/or comparatively large MSE. Some parameters have MSE values < 10^{-2} (Figure 10), but the wide range of uncertainty around these parameters is still indicative of non-identifiability (Figure 9).

Remarkably, R_0 can be precisely estimated with unbiased results for parameter sets $\Theta_1 – \Theta_4$ (Figures 9 & 10). When simultaneously estimating five or more parameters, however, the associated uncertainty of all the parameters results in non-identifiability of R_0. For Θ_5, for example, R_0 estimates vary widely in the range (0.683, 2.821) with an MSE of 0.467. As previously mentioned, R_0 is a threshold parameter (epidemic threshold at R_0=1), so given the confidence interval including the critical value 1, we would not have the ability to distinguish between the potential for epidemic spread versus no outbreak.

3.4. Model 4: Zika Model with human and mosquito populations

For this complex model, we find again that when estimating only 1 or 2 parameters (Θ_1, Θ_2), the parameters can be recovered precisely with unbiased results (Figures 11 & 12). When jointly estimating more than two parameters (Θ_i: i > 2), non-identifiability issues arise. It can be seen that the confidence intervals and MSE for β and γ_{h1} are very small, and thus they are identifiable. However, all of the confidence intervals and MSE values for each of the other parameters (Θ_i: i > 2) are representative of non-identifiability. The parameter estimates have a large amount of uncertainty, represented by the large confidence intervals, and are also biased estimates of the true value: MSE > 10^{-2} for all.

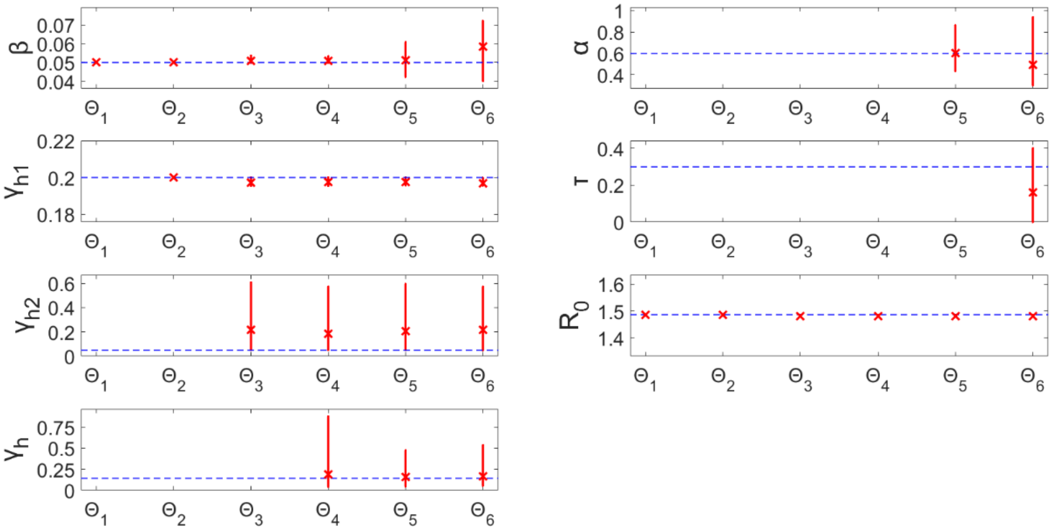


Figure 11. Model 4 – 95% confidence intervals (vertical red lines) for the parameter estimate distributions obtained from the 200 realizations of the simulated datasets. Mean estimated parameter value is denoted by red x, and the true parameter value is represented by the blue horizontal line. Θ_i denotes the estimated parameter set, where i indicates the number of parameters being jointly estimated.

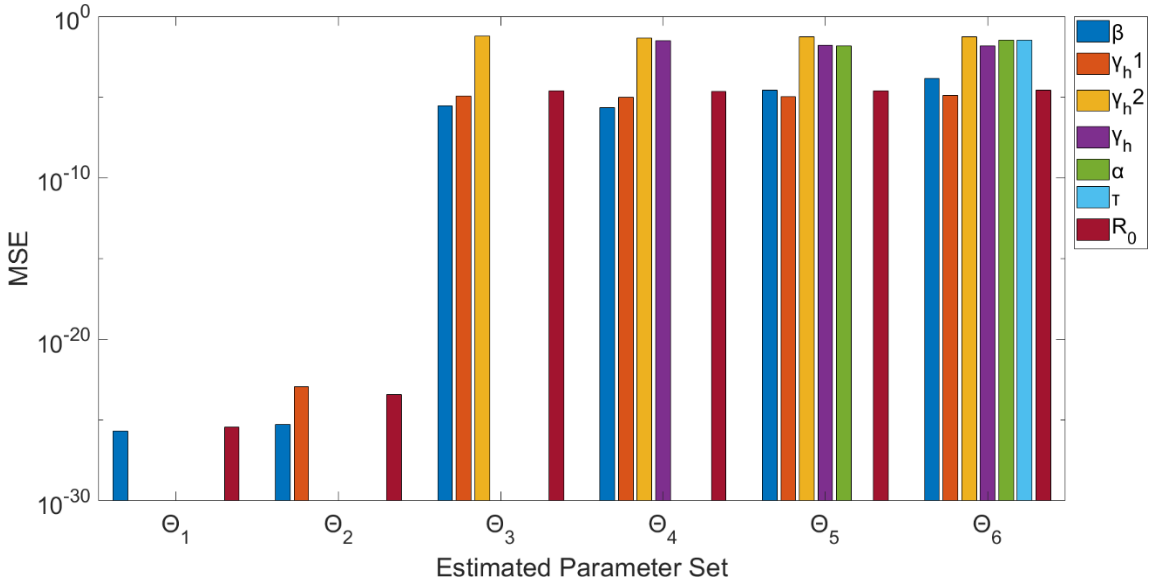


Figure 12. Model 4 – Mean squared error (MSE) of the distribution of parameter estimates (200 realizations) for each estimated parameter set Θ_i, where i indicates the number of parameters being jointly estimated. Note that the y-axis (MSE) is represented with a logarithmic scale.

In terms of R_0, we can see that this composite parameter of interest is identifiable for all Θ_i (Figures 11 & 12). Despite the large confidence intervals associated with some parameters (ex: $\Theta_6 - \gamma_{h2}$: (0.047, 0.573)), when estimating more than two parameters, R_0 can still be estimated with low uncertainty: ($\Theta_6 - R_0$: (1.480, 1.486)). The R_0 estimates have little error, as MSE < 10^{-4} for all Θ_i. This is consistent with the previous models in that R_0 estimates are robust to the uncertainty and bias of the other estimated parameters.

4. Discussion

In this paper we have introduced a simple computational approach for assessing parameter identifiability in compartmental models comprised of systems of ordinary differential equations. We have demonstrated this approach through various examples of compartmental models of infectious disease transmission and control. Using simulated time series of the number of new infectious individuals, we analyzed the identifiability of model characterizing transmission and the natural history of the disease. This type of analysis based on simulated data provides a crucial step in infectious disease modeling, as inferences based on estimates of non-identifiable parameters can lead to incorrect or ineffective public health decisions. Parameter identifiability and uncertainty analyses are essential for assessing the stability of the parameter estimates. Hence, it is important for researchers to be mindful that a good fit to the data does not imply that parameter estimates can be reliably used to evaluate hypotheses regarding transmission mechanisms. Moreover, quantifying the uncertainty surrounding parameter estimates is key when making inferences that guide public health policies or interventions.

Our bootstrap-based approach is sufficiently general to assess identifiability for compartmental modeling applications. We have shown that this method works well for models of varying levels of complexity, ranging from a simple SEIR model with only a few parameters (Model 1) to a complex, dual-population compartmental model with a total of 16 parameters (Model 4). Other methods exist to conduct parameter identifiability analyses. Some methods, such as Taylor series methods (15, 16) and differential algebra-based methods (17, 18), require more mathematical analyses, which becomes increasingly complicated as model complexity increases. Other methods rely on constructing the profile likelihood for each of the estimated parameters to assess local structural identifiability (11, 14, 31, 32). In this method, one of the parameters (θ_i) is fixed across a range of realistic values, and the other parameters are refit to the data using the likelihood function of θ_i. Thus, identifiability of the parameters is determined by the shape of the resulting likelihood profile. Depending on the assumptions of the error structure in the data and as models become increasingly more complex, derivation of the likelihood profile and confidence intervals becomes increasingly more difficult.

Overall, our analyses indicate that parameter identifiability issues are more likely to arise with more complex models (based on number of equations/states and parameters). For example, a set of 3 parameters (Θ_3) can be estimated with low uncertainty and bias from a simple model, like Model 1; however, for more complex models (Model 3, Model 4), estimating only 3 parameters from a single curve of case incidence resulted in lack of identifiability for at least one of the parameters in the set (Θ_3). Also, for Θ_i (recall: i represents number of parameters being jointly estimated), as *i* increases, the uncertainty surrounding estimated parameters tended to increase, on average, as well (Figure 7). One strategy to resolve parameter identifiability issues consists of restricting the number of parameters being jointly estimated while fixing other parameter values and conducting sensitivity analyses.

Importantly, we found that R_0 is a robust composite parameter, even in the presence of identifiability issues affecting individual parameters in the model. In Model 4, despite large confidence intervals and larger MSE for the estimated parameters, R_0 estimates were contained in a finite confidence interval with little bias (Figures 11 & 12). For example, for parameter set Θ_6, only two of the estimated parameters could be reliably identified from the data, yet R_0 could be identified with little uncertainty or bias. These findings are in line with the identifiability results of R_0 for a vector-borne disease model (similar to Model 4), even when other model parameters could not be properly estimated (14). R_0 is often a parameter of interest, as R_0 values have been related to the size or impact of an epidemic (1). Moreover, R_0 estimates can be used to characterize initial transmission potential, assess the risk of an outbreak, and evaluate the impact of potential interventions, so it is beneficial to know we can reliably obtain R_0 estimates, despite lack of identifiability in other parameters.

It is important to emphasize that our methodology is helpful to uncover identifiability issues which could arise from 1) the lack of information in the data or 2) the structure of the model. We also note that our examples assess identifiability of parameters by relying on the entire curve of incidence data of a single epidemic. Future work could include identifiability analyses in the context of limited data using different sections of the trajectory of the outbreak. We also assume that only one model variable (state) is observed, so future analyses could incorporate more than one observed variable to potentially improve the identifiability of parameters without changing the model. For example, for Model 3 (Ebola), the incidence curves of new hospitalized cases and new deaths could provide additional information that better constrain parameter estimates, thereby improving parameter identifiability results.

5. Conclusions

For modeling studies, we recommend conducting comprehensive parameter identifiability analyses based on simulated data prior to attempting to fit the model to data. It is important to emphasize that lack of identifiability could be due to lack of information in the data or the structure of the model. The analyses also help guide the set of parameters in the model that can be jointly estimated – identifiability issues may not arise until any given number of parameters are being simultaneously estimated. If the analysis indicates non-identifiability of certain parameters, may have to be assessed in sensitivity analyses (rather than estimated) to address the identifiability issue.

In summary, the ability to make sound public health decisions regarding an infectious disease outbreak is crucial for the general health and safety of a population. Knowledge of whether a parameter is identifiable from a given model and data is invaluable, as estimates of non-identifiable parameters should not be used to inform public health decisions. Further, parameter estimates should be presented with quantified uncertainty. The methodology presented in this paper adds to the essential toolkit for conducting model-based inferences.

References

1. Anderson RM, May RM. Infectious Diseases of Humans: Dynamics and Control. Oxford: Oxford University Press; 1991.

2. Diekmann O, Heesterbeek JA, Metz JA. On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations. Journal Of Mathematical Biology. 1990;28(4):365-82.

3. Chowell G. Fitting dynamic models to epidemic outbreaks with quantified uncertainty: A primer for parameter uncertainty, identifiability, and forecasts. Infectious Disease Modelling. 2017;2:379-98.

4. He D, King A, King AA, Ionides EL. Plug-and-play inference for disease dynamics: measles in large and small populations as a case study. Journal of the Royal Society Interface. 2010;7(43):271-83.

5. Goeyvaerts N, Willem L, Van Kerckhove K, Vandendijck Y, Hanquet G, Beutels P, et al. Estimating dynamic transmission model parameters for seasonal influenza by fitting to age and season-specific influenza-like illness incidence. Epidemics. 2015;13:1-9.

6. Chowell G, Viboud C, Simonsen L, Merler S, Vespignani A. Perspectives on model forecasts of the 2014-2015 Ebola epidemic in West Africa: lessons and the way forward. BMC Medicine. 2017;15(1):42-.

7. Banks HT, Holm K, Robbins D. Standard error computations for uncertainty quantification in inverse problems: Asymptotic theory vs. bootstrapping. Mathematical and Computer Modelling. 2010;52:1610-25.

8. Gibson GJ, Streftaris G, Thong D. Comparison and assessment of epidemic models. Statistical Science. 2018;33(1):19-33.

9. Banks H, Davidian M, Samuels Jr J, Sutton K. An inverse problem statistical methodology summary. In: Chowell G, Hyman J, Bettencourt L, Castillo-Chavez C, editors. Mathematical and statistical estimation approaches in epidemiology. Dordecht, The Netherlands: Springer; 2009. p. 249-302.

10. Wu KM, Riley S. Estimation of the Basic reproductive number and mean serial interval of a novel pathogen in a small, well-observed discrete population. PLoS ONE. 2016;11(2):1-12.

11. Breto C. Modeling and inference for infectious disease dynamics: A likelihood-based approach. Statistical Science. 2018;33(1):57-69.

12. Scranton K, Knape J, de Valpine P. An approximate Bayesian computation approach to parameter estimation in a stochastic stage-structured population model. Ecology. 2014(5):1418.

13. Abdessalem AB, Dervilis N, Wagg D, Worden K. Model selection and parameter estimation in structural dynamics using approximate Bayesian computation. Mechanical Systems and Signal Processing. 2018;99:306-25.

14. Kao Y-H, Eisenberg M. Practical unidentifiability of a simple vector-borne model: implications for parameter estimation and intervention assessment. Epidemics. 2018.

15. Miao H, Xia X, Perelson AS, Wu H. On identifiability of nonlinear ODE models and applications in viral dynamics. SIAM Review. 2011(1):3.

16. Pohjanpalo H. System identifiability based on power-series expansion of solution. Mathematical Biosciences. 1978(41):21-33.

17. Eisenberg MC, Robertson SL, Tien JH. Identifiability and estimation of multiple transmission pathways in cholera and waterborne disease. Journal of Theoretical Biology. 2013;324:84-102.

18. Ljung L, Glad T. Testing global identifiability for arbitrary model parameterizations. IFAC Proceedings Volumes. 1991;24:1085-90.

19. Chis O-T, Banga JR, Balsa-Canto E. Structural identifiability of systems biology models: A critical comparison of methods. PLoS ONE. 2011;6(11):1-16.

20. Lloyd A. Introduction to Epidemiological Modeling: Basic Models and Their Properties; 2007.

21. Brauer F, van der Driessche P, Wu J, Allen LJS. Mathematical Epidemiology. Berlin: Springer; 2008.

22. van den Driessche P, Watmough J. Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission. Mathematical Biosciences. 2002;180:29-48.

23. Chowell G, Nishiura H. Comparative estimation of the reproduction number for pandemic influenza from daily case notification data. Journal of the Royal Society Interface. 2007;4(12):155-66.

24. Chowell G, Ammon CE, Hengartner NW, Hyman JM. Estimation of the reproductive number of the Spanish flu epidemic in Geneva, Switzerland. Vaccine. 2006;24:6747-50.

25. Legrand J, Grais RF, Boelle PY, Valleron AJ, Flahault A. Understanding the dynamics of Ebola epidemics. Epidemiology and Infection. 2007(4):610.

26. Gao D, Lou Y, He D, Porco TC, Kuang Y, Chowell G, et al. Prevention and control of Zika as a mosquito-borne and sexually transmitted disease: A mathematical modeling analysis. Scientific Reports. 2016;6:28070-.

27. Efron B, Tibshirani R. An introduction to the bootstrap. New York: Chapman & Hall; 1993.

28. Chowell G, Hengartner NW, Castillo-Chavez C, Fenimore PW, Hyman JM. The basic reproductive number of Ebola and the effects of public health measures: The cases of Congo and Uganda. 2005.

29. Cobelli C, Romanin-Jacur G. Controllability, observability and structural identifiability of multi input and multi output biological compartmental systems. IEEE Transactions on Biomedical Engineering. 1976;BME-23(2):93.

30. Jacquez JA. Compartmental analysis in biology and medicine. Ann Arbor: University of Michigan Press, c1985. 2nd ed.; 1985.

31. Raue A, Kreutz C, Maiwald T, Bachmann J, Schilling M, Klingmuller U, et al. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. Bioinformatics. 2009;25(15):1923-9.

32. Nguyen VK, Binder SC, Boianelli A, Meyer-Hermann M, Hernandez-Vargas EA. Ebola virus infection modeling and identifiability problems. Frontiers in Microbiology. 2015;6.

Chapter 3. Comparative assessment of parameter estimation methods in the presence of overdispersion: a simulation study

Roosa, K., Luo, R., & Chowell, G. (2019) Comparative assessment of parameter estimation methods in the presence of overdispersion: a simulation study. *Mathematical Biosciences and Engineering.* doi: 10.3934/mbe.2019214

1. Background

Mathematical modeling offers a quantitative framework to investigate dynamics of infectious disease epidemics and guide public health decisions regarding the type and intensity of control interventions. The mathematical modeling toolkit includes phenomenological models that assess features in epidemic trajectories and mechanistic models with which researchers strive to evaluate the effects of interventions or the roles of different factors on transmission dynamics (e.g., mixing patterns or environmental factors). Dynamic models based on differential equations are often calibrated using infectious disease outbreak data that typically correspond to time series of new cases, where a "case" corresponds to an observable (reportable) event. Further, the dataset corresponds to only one realization of a stochastic process and, unfortunately, in the context of real outbreaks occurring in natural environments, generating more data realizations in a carefully controlled environment is not feasible.

When calibrating models to data via some fitting process (also known as data assimilation), the solution of the dynamic model for a given set of parameter values and initial conditions is typically considered to be the "mean" solution, which is embedded into a counting process characterized by a statistical model (e.g. Poisson, Negative Binomial). For instance, a researcher fits an SEIR-type model ("mean" signal) to weekly series of newly reported cases of Ebola in West Africa assuming a Poisson error structure around the case series data. In this inference framework, the *equidispersion* property of the Poisson distribution (where the mean is equal to the variance) simplifies the inference process, limits the number of degrees of freedom, and indirectly reduces potential issues of parameter non-identifiability [1]. Moreover, in real-time analyses of evolving outbreaks, one has to link the observable with the unobservable by adjusting, for instance, for delays associated with incubation periods, time to diagnosis, or reporting delays.

Importantly, visual inspection of time series data could suggest an apparently larger variability than the mean signal linked to the model. One potential source of this *apparent overdispersion effect* could arise from systematic deviations of the model ("mean") to the data due to model misspecification (e.g., the model incorrectly specifies the length of the incubation period or neglects another important mechanism involved in the dynamic process) [2]. Hence, researchers could fix this lack of model fit by identifying and incorporating other key process components in the model, thus resolving the apparent overdispersion issue. Alternatively, there is actual *overdispersion,* where the variability in the data is larger than expected. In this case, the researcher may reconsider the statistical model employed to model the error structure in the data by considering error structures that allow for the variance to be larger the mean (e.g., Negative Binomial) [1]. Hence, identifying the relevant sources of apparent overdispersion is critical in the modeling process as it could lead to poor descriptions of the data and predictive power and underestimated standard errors and confidence intervals [3].

Fortunately, simulation studies can be utilized to evaluate the impact of various forms of misspecification when calibrating a model to data. As explained above, such modeling challenges could be related to the model design or to the variability in the data. In a previous paper, we outline a simple computational bootstrap-based method for assessing parameter identifiability [4]. This method involves repeated sampling from the deterministic model solution to simulate multiple data sets from which the parameters are re-estimated, which allows us to detect parameter non-identifiability that could arise from model structure or the amount of information that can be extracted from the available data. This method was originally devised to quantify parameter uncertainty [4-6]. In this paper, we evaluate the effects of misspecification of the error structure (in the data) on bias and uncertainty associated with parameter estimates using simple dynamic transmission models. Specifically, we focus on modeling varying levels of data overdispersion stemming from randomness in the counting process that shapes the time series data, rather than systematic misspecifications in the mean process linked to the dynamic model. We utilize the parametric bootstrap approach to assess parameter estimates and their uncertainty as a function of the level of random noise in the data, and we compare results using two common parameter estimation methods: nonlinear least squares (LSQ) and maximum likelihood estimation with a Poisson error structure (Poisson-MLE).

2. Methods

2.1. Phenomenological models

For each of the following model examples, daily time series incidence (total number of new cases) curves were simulated directly from the model equation. All simulations and analyses were performed in Matlab 2017 (Mathworks, Inc).

Example 1: Generalized growth model (GGM)

Models used to study the growth patterns of infectious disease outbreaks often assume exponential growth in the absence of control interventions (compartmental models, for example) [7, 8]; however, growth patterns are likely slower than exponential for some diseases depending on the mode of transmission and the population structure. For example, Ebola spreads only via close contact, so in a constrained population contact structure, sub-exponential growth patterns would be expected [9]. The generalized growth model (GGM) includes a “deceleration of growth” parameter, also referred to as a “scaling of growth” parameter, p (range: [0, 1]) that relaxes the assumption of exponential growth [10]. A value of p = 0 represents constant (linear) growth, while a value of p = 1 indicates exponential growth. If 0 < p < 1, the growth pattern is characterized as sub-exponential or polynomial.

The GGM is as follows:

$$\frac{dC(t)}{dt} = \dot{C}(t) = rC(t)^{p},$$

where C(t) describes the cumulative number of cases at time t, $\dot{C}(t)$ is the incidence curve, r is the growth rate parameter (r > 0), and p is the deceleration of growth parameter [10].

Allowing for a range of growth scaling in the model allows for applications to outbreak data for various different diseases. For example, the GGM has been applied to forecast outbreaks of a range of diseases, including foot and mouth disease [11], Zika [12], pandemic influenza [13], HIV/AIDS [14], and Ebola [15, 16]. For the example presented here, we assume a growth rate r = 0.4 and a deceleration of growth rate p = 0.9 (Table 1) for the simulated data. Similar values have been used to characterize pandemic influenza. Additional examples analyzing other parameter set values are provided in the supplemental material.

Table 1. Generalized growth model

Parameters	**Description**	**Value**	**Bounds**
r	Rate of change (growth rate)	0.4	[0, 10]
p	Deceleration constant (0≤p≤1)	0.9	[0, 1]

Example 2: Generalized logistic growth model (GLM)

While the GGM can model early epidemic growth, the function is strictly increasing, and thus cannot be used to fit entire epidemic curves (as it is assumed epidemic growth will slow at some point in time). The generalized logistic growth model (GLM) is an extension of the GGM that includes a parameter K that classifies the carrying capacity or final size of the epidemic. The GLM is as follows:

$$\dot{C}(t) = rC(t)^p(1 - \frac{C(t)}{K}),$$

where C(t) describes the cumulative number of incident cases at time t, and $\dot{C}(t)$ is the incidence curve [10]. Again, r is the intrinsic growth rate, p is the deceleration of growth parameter, and K is the final epidemic size. For the GLM, we use the same r and p values from *Example 1* and set the final epidemic size K = 10,000 (Table 2).

Table 2. Generalized logistic growth model

Parameters	**Description**	**Value**	**Bounds**
r	Rate of change (growth rate)	0.4	[0, 10]
p	Deceleration constant (0≤p≤1)	0.9	[0, 1]
K	Final epidemic size/ carrying capacity	10000	[0, 1000000]

2.2. Data error structure

The Poisson distribution is a commonly assumed error structure for count data, as it has equality of the mean and variance [1, 3]. Empirically, data often represent overdispersion, where the observed variance is higher than the assumed model variance, which may be explained by model misspecification or missing crucial information about the disease [17]. If the mechanism producing the overdispersion is known, it could be remedied by revising the model; however, when it is unknown, it is typically assumed that the variance in the data exceeds the mean (by some scaling factor) [1, 6]. Relative to the Poisson distribution, the negative binomial distribution requires an additional parameter to model count data with varying levels of overdispersion. In the case of equidispersion (variance equal to the mean) the Poisson distribution is a special case of the negative binomial distribution.

For the simple phenomenological models employed here, the data are realizations of random counts $y_{t_i} = y_{t_1}, y_{t_2}, \dots, y_{t_n}$ (i = 1, 2, …, n) following the defined distribution, where t_i are time points and *n* is the number of observations [6]. We model the error using the negative binomial distribution with E(Y) = μ and $var(Y) = \sigma^2$. Let $d = \sigma^2/$μ represent the variance-to-mean ratio. Thus d = 1 yields the Poisson distribution, and d > 1 represents cases of overdispersion, for which the negative binomial distribution is a common choice [1]. It should be noted that d < 1 represents underdispersion, though this is rarely seen in empirical data [2]. Here, we consider data simulated under the following values: *d* = [1, 2, 20, 40, 60, 80, 100], which represent different levels of overdispersion.

2.3. Simulated data

We utilize a Monte Carlo simulation to quantify the uncertainty of parameter estimates. This approach is an iterative process that involves simulating random error around the data in each iteration to generate a sample of data sets from which to estimate parameters [18]. The estimated parameters from all iterations simulate the sampling distributions from which we construct the confidence intervals of parameters. This method is similar to the parametric bootstrapping approach (derived from the general bootstrap method [6]), which first fits the model of interest to the available time-series data to obtain the best-fit estimate of the parameters; however, here we are simulating the data directly from the model, so the ’best-fit estimate’ to the simulated data is essentially the model with the given parameter values (Θ_{true}). This way we know the true parameter values of the data and can assess the performance of different parameter estimation approaches considering different error structures.

We generate time series data directly from the model equation, setting parameters to values of interest (Tables 1 and 2); this yields the data y_t with solution $f(\text{t}, \Theta_{true})$ which is used to generate M = 500 simulated data sets for each variance-to-mean ratio. Assuming Poisson error structure, each new observation is sampled from a Poisson distribution with mean = $f(\text{t}, \Theta_{true})$. which is the incidence curve, at each time point t. This results in 500 estimated parameter sets $\hat{\Theta}_i$, where i = 1, 2, …, M.

To analyze scenarios of data overdispersion, or of variance greater than Poisson mean, we utilize the negative binomial distribution with variance-to-mean ratios (d) of 2, 20, 40, 60, 80, and 100. To simulate negative binomial noise, the variance at each time point t is given by multiplying $f(\text{t}, \Theta_{true})$, the mean, by the specified variance-to-mean ratio. The above steps (1-4) are repeated for each value of d, assuming a negative binomial error structure, in place of Poisson. Thus, we obtain empirical distributions for each estimated parameter at each of the seven variance-to-mean ratios, where d = 1 indicates Poisson noise.

2.3. Parameter Estimation

For each example, we run the bootstrapping analyses using the two general estimation methods. For both estimation methods, one can use numerical optimization methods available in Matlab or R (R Core Team). The methods are as follows:

Method 1: Nonlinear least squares (LSQ)

Least squares estimation yields the best fit solution to the time series data by searching for the parameter set $\hat{\Theta}$= ($\hat{\theta}_1$, $\hat{\theta}_2$,…, $\hat{\theta}_{\text{m}}$) that minimizes the sum of squared deviations between the data y_t and the corresponding model solution $f(t, \Theta)$ [6]. That is,

$$\hat{\Theta} = argmin \sum_{t=1}^{n} (f(t; \Theta) - y_t)^2.$$

For the presented examples, $f(t, \Theta) = C'(t|\Theta)$ is the function of the incidence rate at time *t* that depends on the set of parameters Θ. For the GGM (Example 1), Θ = (*r*, *p*). For the GLM (Example 2), Θ = (*r, p, K*). Then, $\hat{\Theta}$ is the parameter set that yields the smallest differences between the data and model. To run nonlinear least squares estimation (LSQ), we utilize the *fmincon* function in Matlab 2017, which finds the minimum of a constrained nonlinear multivariable function, with the 'interior-point' algorithm (default). Further, we restrict the bounds for the parameters (Tables 1 & 2).

This parameter estimation method gives the same weight to all of the data points. LSQ also does not require a specific distributional assumption for y_t, except for the first moment $E[y_t] = f(t_i; \Theta)$; meaning, the mean at time *t* is equivalent to the count (e.g., number of cases) at time *t* [19]. LSQ fitting yields asymptotically unbiased point estimates regardless of any misspecification of the variance-covariance error structure. It is of interest to study the impact of data overdispersion on LSQ parameter estimates.

Method 2: Poisson-MLE

The goal of maximum likelihood estimation (MLE) is to derive parameter estimates given a model that dictates the dynamical process and data with variability assumed to follow a specific probability distribution. Consider the probability density function (PDF) that specifies the probability of observing data y_t given the parameter set Θ, or $f(y_t|\Theta)$; given a set of parameter values, the PDF can show which data are more probable, or more likely [19]. MLE aims to determine the values of the parameter set that maximizes the likelihood function, where the likelihood function is defined as $L(\Theta|y_t) = f(y_t|\Theta)$ [19, 20]. The resulting parameter set is called the MLE estimate, the most likely to have generated the observed data. Specifically, the MLE estimate is obtained by maximizing the corresponding log-likelihood function. For count data with variability characterized by the Poisson distribution, we utilize Poisson-MLE, where the log-likelihood function is given by:

$$L(\Theta|y_{t_i}) = \sum_{i=1}^{n} [y_{t_i}\, log(f(t_i; \Theta)) - f(t_i; \Theta)]$$

and the Poisson-MLE estimate is expressed as

$$\hat{\Theta} = argmax \sum_{i=1}^{n} [y_{t_i} \, log(f(t_i;\Theta)) - f(t_i;\Theta)].$$

We again utilize the *fmincon* function with the same parameter bounds as defined for LSQ (Tables 1 & 2). It is of interest to compare the performance of Poisson-MLE and LSQ with simulation in the context of increasing levels of variability in the data.

We utilize Poisson-MLE and LSQ to estimate the parameters for each simulated data set, resulting in 500 estimated parameter sets $\hat{\Theta}_{\mathrm{I}}$ where i = 1, 2, …, M (per level of overdispersion). We repeat the steps for each specified level of overdispersion, with variance-to-mean ratios given by d = 1 (Poisson), 2, 20, 40, 60 , 80, and 100.

2.5. Performance

To compare the parameter estimation methods within each model example, we assess the empirical distributions of the estimates obtained from Monte Carlo simulation. For each method, we calculate the 95% confidence intervals (CIs) for each parameter using the 2.5 and 97.5 percentiles. The width of the confidence intervals is used to compare the uncertainty of parameter estimates at each level of overdispersion. As LSQ has wider confidence intervals in the majority of the simulations (with only one exception), the relative width differences between two confidence intervals is calculated as: $\% \, diff = \frac{width_{LSQ} - width_{MLE}}{width_{LSQ}} \times 100\%$.

Further, we use the mean squared error (MSE) to quantify accuracy, or how close the estimated values are from the true parameter value across the entire distribution of parameter estimates. MSE is calculated as: $MSE = \frac{1}{M} \sum_{i=1}^{M} (\theta_{true} - \hat{\theta}_i)^2$, where θ_{true} represents the true parameter value (in the simulated data) and $\hat{\theta}_i$ represents the estimated parameter value for the ith bootstrap sample.

3. Results

Example 1: Generalized growth model (GGM)

For the GGM, we simultaneously estimate both model parameters, r and p, and compare results for the two estimation methods: nonlinear least squares (LSQ) and maximum likelihood estimation (MLE). We use an ascending phase length (amount of data fit to) of 45 days, and we will later assess how the amount of data used impacts the results. Based on both LSQ and Poisson-MLE, the level of overdispersion in the data had little effect on the mean estimated parameter values. This can be seen in Figure 1, as the mean estimates at each level are distributed randomly around and very closely to the true value line. For both methods (LSQ and Poisson-MLE), the amount of uncertainty surrounding the parameter estimates increases as the level of overdispersion increases – 95% confidence intervals become increasingly wider for higher variance-to-mean ratios (Figure 1).

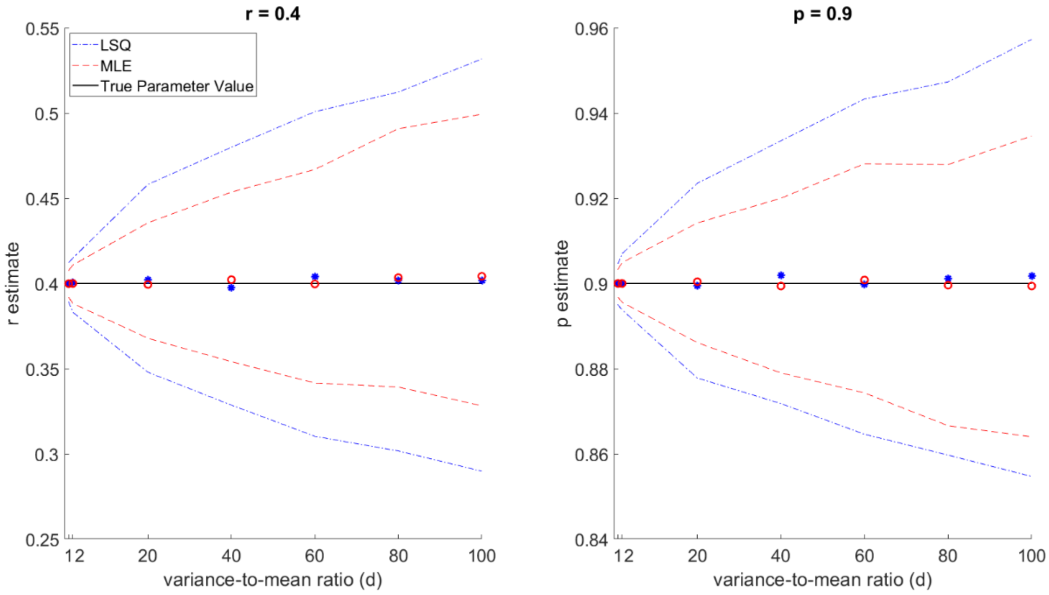


Figure 1. GGM parameter estimation results for increasing levels of variance assumed in the data. Mean estimates (circles) and 95% confidence intervals (dashed lines) are shown for the two estimation methods: nonlinear least squares (LSQ) and maximum likelihood estimation (MLE).

The percent differences of CI widths (for each *d*) are nearly equivalent for both r and p, so to avoid repetition, we will discuss percent difference in terms of r. Results indicate that even small levels of overdispersion can impact the uncertainty; for MLE, there is a 43% increase in CI width from *d* = 1 (r = 0.4 (0.392, 0.408)) to *d* = 2 (r = 0.4 (0.388, 0.411)), and for LSQ a 38% increase from *d* = 1 (r = 0.4 (0.389, 0.412)) to *d* = 2 (r = 0.4 (0.383, 0.415)) is observed. The largest increase in CI width is seen between *d* = 2 (see CIs in previous sentence) to *d* = 20 (MLE: r = 0.399 (0.368, 0.436); LSQ: r = 0.402 (0.348, 0.458)), with a 201.33% increase for MLE and a 246.54% increase for LSQ. After *d* = 20, the variance-to-mean ratio increases by 20. For each of these increases in variance-to-mean ratio, the % increase in CI width ranges from 13% - 47% for MLE and 11% - 37% for LSQ.

Comparing the methods to each other, we can see that MLE consistently yields narrower confidence intervals, or less uncertainty, compared to LSQ. Across the levels of overdispersion, the relative difference between LSQ and MLE confidence interval widths ranges from 28-38.5% for r and 29.5-38.5% for p, showing that while both methods' CIs are increasing, the relative difference between them is remaining stable. While these relative differences may seem high, actual CI width differences between the methods are small. With Poisson data, a difference in CI width of 0.0073 is observed for r and 0.0031 for p, comparing LSQ to MLE. For a large amount of overdispersion (d = 100), the CI difference is 0.0708 (29.27%) for r and 0.032 (31.19%) for p. In regards to application of the GGM specifically, these differences in widths do not yield practically meaningful differences in parameters estimated.

In Figure 2, we see that the MSE increases for higher levels of noise in the data. At each level of overdispersion, MLE and LSQ have very small differences in MSE, but MLE consistently has lower MSE. The MSE for LSQ ranges from 2.05-2.38 times the MSE for MLE for r and 2.01-2.30 times for p. Again, while MLE is relatively more accurate than LSQ, the practical differences are small. For example, the largest difference in MSE seen between the two methods was less than 0.002.

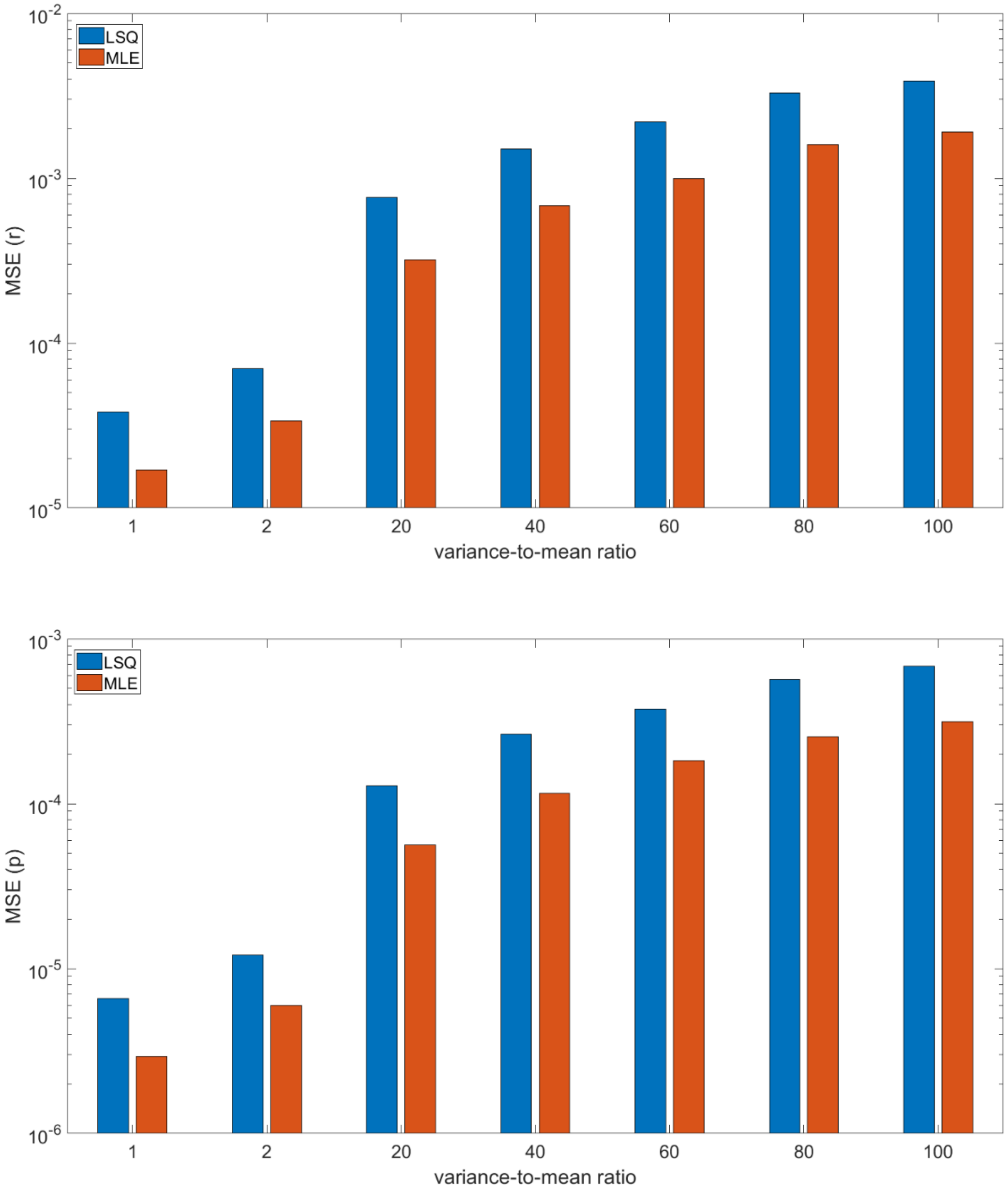


Figure 2. Mean squared error of the distribution of GGM parameter estimates (M=500) for increasing levels of error assumed (variance-to-mean ratio) is shown for both estimation methods: nonlinear least squares (LSQ) and maximum likelihood estimation (MLE). Results for r are shown in (a), and results for p are in (b). Note that MSE is in log-scale.

3.2. Example 2: Generalized logistic growth model

For the generalized logistic model (LGM), we jointly estimate all three parameters: r, p, and K. Again, we see that the uncertainty surrounding the parameters (width of the confidence intervals) increases with the increasing variance-to-mean ratio (Figure 5). Again, the largest % difference observed was from $\sigma^2 = 2$ to $\sigma^2 = 20$, with increases in CI width of 232%, 226%, and 264% (r, p, and K, respectively) for MLE, and 235%, 235%, and 272% for LSQ. % increases were comparable for the two methods, with LSQ consistently yielding slightly wider CIs for r and p, and nearly equivalent CIs for K (Figure 5).

For both LSQ and MLE, the mean parameter estimates remain around the true value line for r and K. The estimates for p begin to deviate from the true line in an upward trajectory, with the mean estimates from LSQ rising faster (so further from the true value) than MLE. These differences in bias, however, are again very small, with the largest difference in MSE for r and p being less than 0.0025. Further, for K estimates, MLE and LSQ yielded minimal differences in MSE, or bias (Figure 6).

It is known that parameter estimation results depend on the amount of information in the data available to fit the model to. For this purpose, we also performed the analyses for four different variance-to-mean ratios (d = 1, 10, 50, 100) at increasing lengths of the ascending phase, ranging from 15 to 40 days (increments of 5 days). Across each level of overdispersion, the overall pattern was consistent: fitting to more data (longer ascending phase length) resulted in smaller confidence intervals, and thus, lower uncertainty of parameter estimates. While the widths of the confidence intervals vary significantly across the levels of overdispersion, this general pattern is clearly seen for both estimation methods (Figure 3).

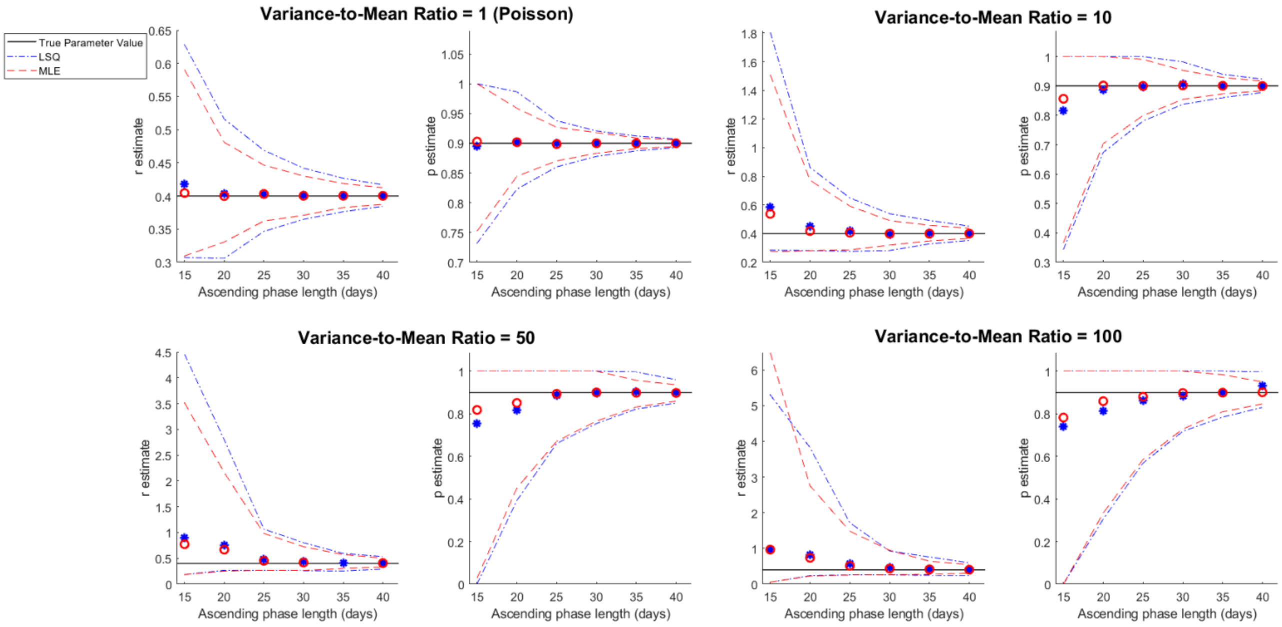


Figure 3. GGM parameter estimates as amount of data available (length of ascending phase) increases. Variance-to-mean ratios of: 1, 10, 50, 100. Mean estimates are represented by the circles and 95% confidence intervals are represented with dashed lines.

For the Poisson error structure (*d* = 1), Poisson-MLE should yield the 'true' uncertainty (confidence intervals) of estimates, as the method is based on the correct error structure; whereas, LSQ assumes constant variance. It is shown that, for the Poisson case, LSQ yields wider confidence intervals compared to MLE (CI width differences - LSQ - MLE - for r: 0.040, 0.060, 0.038, 0.018, 0.014, 0.007 for ascending phase length, t = 15, 20, 25, 30, 35, and 40 days respectively), and the difference between the two methods decreases as the ascending phase increases, excluding from 15 to 20 days (Figure 3). The % difference in width between the two methods ranges from 12.43 - 31.07% (for all t), but as previously stated, the practical significance of these differences is small.

Figure 3 also shows that higher levels of overdispersion require longer ascending phase lengths to reduce parameter uncertainty. For example, under Poisson assumptions, MLE yields a confidence interval length of about 0.15 for estimates of r using an ascending length of 20 days. When overdispersion is present, longer ascending phases are needed for MLE to yield comparable confidence intervals. Comparing to the CI width of 0.15 for r for MLE with 20 days of data, a variance-to-mean ratio of 10 yields similar uncertainty (MLE CI width: 0.17) with 30 days of data; and further, a variance-to-mean ratio of 50 yields similar uncertainty (MLE CI width: 0.17) with 40 days of data. This indicates that uncertainty of parameter estimates in the presence of overdispersion can be mitigated with the inclusion of more data points. This idea is seen in Figure 3, in that the confidence bounds quickly converge to the true value line as more data points are used, even for extreme cases of overdispersion (e.g., variance-to-mean ratio = 100).

Similarly, at each level of overdispersion, the MSE decreases for longer ascending phase lengths (Figure 4), indicating that more data yields higher accuracy of parameter estimates. For example, for data with Poisson error structure, each 5 day increase in ascending phase yielded between 55% - 72% decrease, in descending order, for r estimated with MLE. Poisson-MLE results for p and LSQ results (r and p) are nearly equivalent and follow the same patterns. It can be seen that, for increasing variance-to-mean ratios, more data points are required before the mean estimate falls on the true value line (Figure 3). For example, looking at the plots of r estimates, the mean value falls on the line around 20 days, 25 days, 30 days, and 35 days for the increasing levels of overdispersion (σ^2 = 1, 10, 50, 100), indicating that these would be the minimum amount of data from which the signal can be accurately detected.

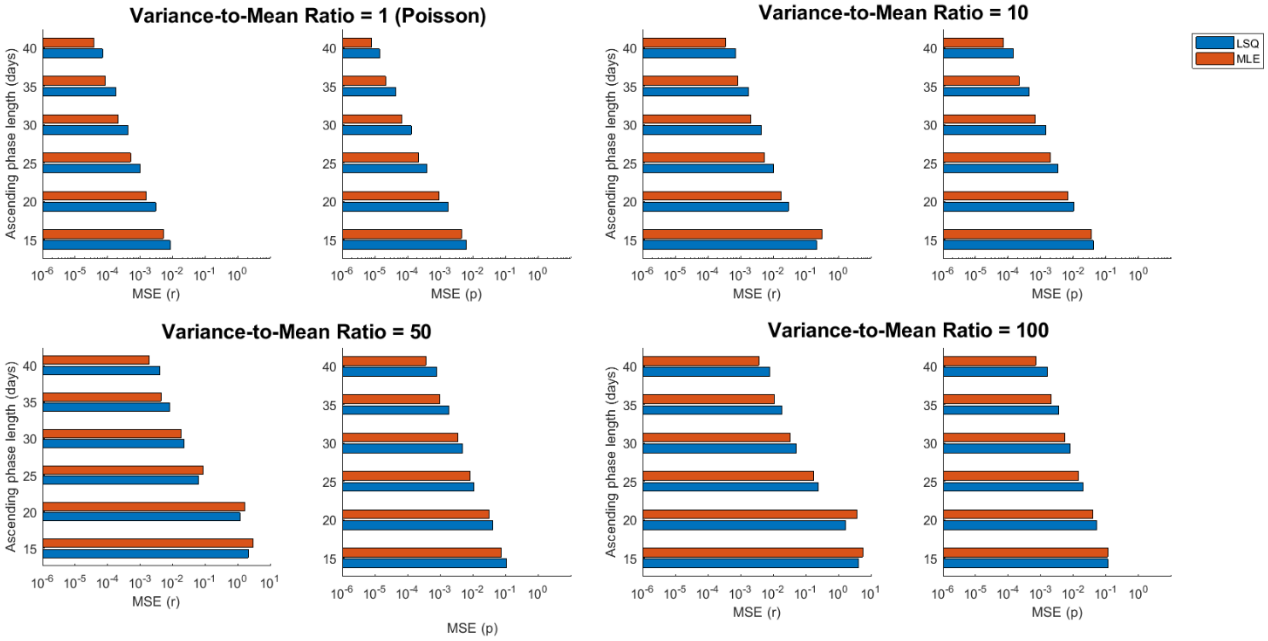


Figure 4. MSE of GGM parameter estimates for each estimation method (LSQ, MLE) as amount of data available (length of ascending phase) increases. Variance-to-mean ratios of: 1, 10, 50, 100.

Example 2: GLM

For the generalized logistic growth model (GLM), we jointly estimate all three parameters: r, p, and K. Again, we see that the uncertainty surrounding the parameters (width of the confidence intervals) increases with the increasing variance-to-mean ratio (Figure 5). Again, the largest % difference observed was from *d* = 2 to *d* = 20, with increases in CI width of 232%; 226%, and 264% (r, p, and K, respectively) for MLE, and 235%, 235%, and 272% for LSQ. Percent increases were comparable for the two methods, with LSQ consistently yielding slightly wider CIs for r and p, and nearly equivalent CIs for K (Figure 5).

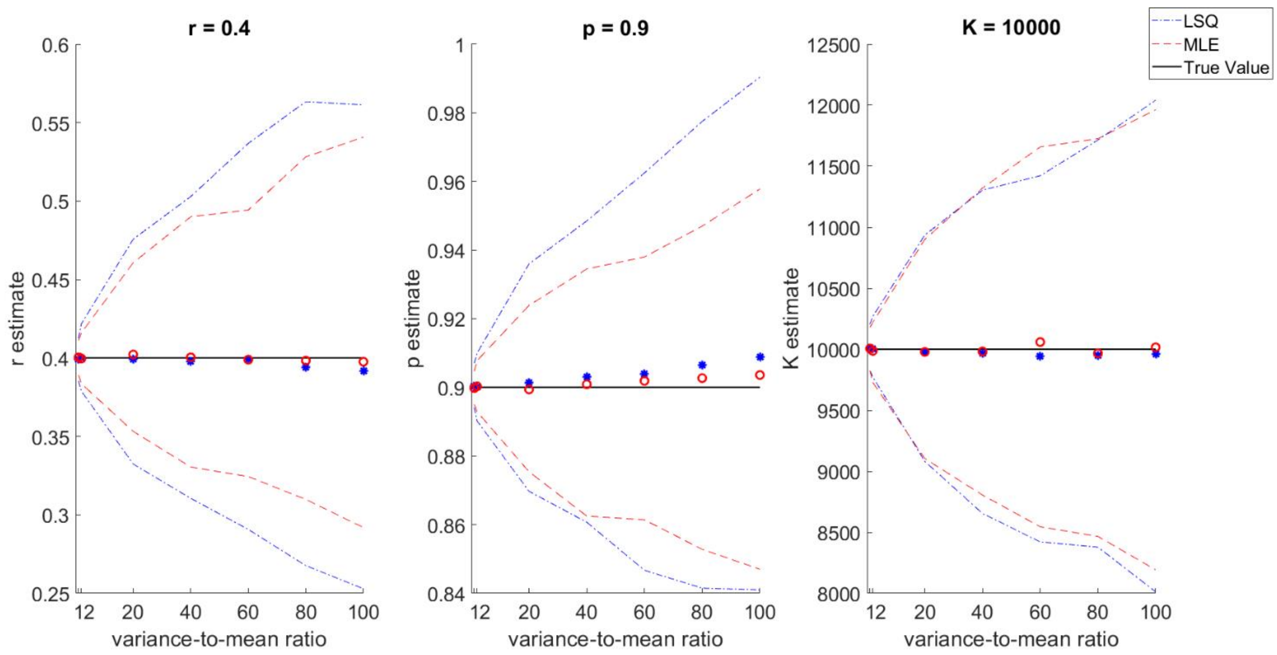


Figure 5. GLM parameter estimation results for increasing levels of variance assumed in the data. Mean estimates (circles) and 95% confidence intervals (dashed lines) are shown for the two estimation methods: nonlinear least squares (LSQ) and maximum likelihood estimation (MLE).

Comparing the methods, we find similar results for the uncertainty of r and p as with the GGM. LSQ consistently yields CIs that are 17.1-33.8% wider for r and 18-33.9% wider for p. Again, these CI width differences are small for the parameters (<0.08 for all), but yield high relative differences due to the small values. For parameters of larger magnitude, like K, the relative difference between the methods is much smaller. The relative CI width difference comparing LSQ to MLE ranges from -3.7 - 6.8% across the levels of overdispersion. At *d* = 60, LSQ has a smaller CI width, hence the negative 3.7%.

For both LSQ and MLE, the mean parameter estimates remain around the true value line for r and K. The estimates for p begin to deviate from the true line in an upward trajectory, with the mean estimates from LSQ rising faster (so further from the true value) than MLE. The MSE for LSQ ranges from 1.4-2.05 times and 1.46-2.25 times the MSE for MLE (r and p, respectively). These differences are again very small, with the largest difference in MSE for r and p being less than 0.0025. Further, for K estimates, MLE and LSQ yielded minimal differences in MSE, or accuracy (Figure 6). For K, the MSE for LSQ ranged from 0.84-1.08 times the MSE for MLE. This indicates that the MSE is at times larger for LSQ and at times larger for MLE; for *d* = 1; 2; 40; and 60, the MSE is smaller for LSQ, and for *d* = 20; 80; 100, the MSE is lower for MLE. This may indicate that LSQ can provide more accurate estimates than MLE for data with lower levels of overdispersion, but MLE provides more accurate estimates for highly overdispersed data.

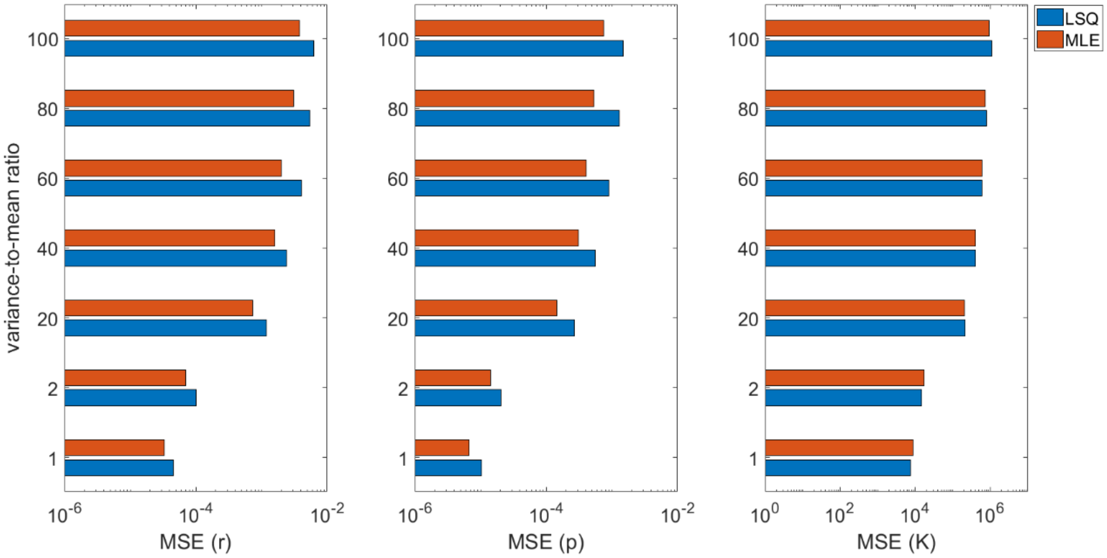


Figure 6. Mean squared error of the distribution of GLM parameter estimates (M=500) for increasing levels of error assumed (variance-to-mean ratio) is shown for both estimation methods: nonlinear least squares (LSQ) and maximum likelihood estimation (MLE). Results for r, p, and K are presented from left to right.

4. Discussion and Conclusions

The results of the uncertainty analyses show a clear pattern of increasing uncertainty (assessed by CI width) as the variance-to-mean ratio (*d*), or overdispersion, increases. The examples (Figures 1 & 5) show near linear growth for the upper and lower bounds of the estimates, increasing as overdispersion increases. Results for r and p for both GGM and GLM reveal that MLE consistently yields more accurate (lower MSE) and precise (smaller CI width) estimates compared to LSQ, though it should be noted that practical differences in estimates are small. For the large parameter, K, differences between methods are even smaller. Across all levels of overdispersion, the relative difference in CI width is less than 6.9%. Further, neither of the methods consistently yields lower MSE for K, indicating comparable MSE, or accuracy, of the methods for large-scale parameters.

For the generalized logistic model (Example 2), we used the entire incidence curve to fit the model, but as explained above, the GGM (Example 1) is strictly increasing and is not flexible enough to detect a peak or decline phase. To illustrate how the amount of data affects the results of parameter estimation for the GGM, we used an increasing number of days (ranging from 15-40 days) for the ascending phase and conducted the analysis across the range for four difference levels of overdispersion. It is clearly seen (Figure 3) that using more data for the ascending phase decreases the uncertainty of parameter estimates, for all levels of overdispersion in the data.

Not only do the confidence interval widths significantly decrease as the ascending phase increases, but the mean estimates trend toward the true value, resulting in smaller MSE values as well (Figure 3, 4). In terms of bias, our simulation study indicates that using too few data points in the presence of overdispersion may result in biased parameter estimates. Each 5 day increase in ascending phase length yielded an improvement in MSE. Further, both estimation methods are based on unbiased estimating equations, and thus there is essentially no difference in terms of bias between the two methods. Figure 4 shows minimal differences in MSE between the two methods for each ascending phase length within each variance-to-mean ratio.

The general descending pattern of the percent improvement in MSE (for 5 day increases in ascending phase) is mostly consistent for each level of overdispersion, though at variance-to-mean ratios of 50 and 100, this pattern is not seen until t = 25. This suggests that the signal still cannot be distinguished from the noise for ascending phases lower than 25 days, and thus an increase of 5 days does not provide much improvement in model fit. While significantly more data is required in the presence of overdispersion, the results suggest that even when overdispersion is suspected, uncertainty and bias of estimation results can be mitigated as more data become available.

These analyses were conducted using data simulated directly from the model corresponding to each example. This is a limitation in that we cannot generalize these results to scenarios with real-world data issues. Further, the results are specific to the parameter values specified, and thus cannot be generalized to all configurations of the models. Because of the difference in results between small parameters (r; p) and large parameters (K) in this study, future studies should look into this pattern (MLE slightly outperforms LSQ for small parameters but performs equivalently for large parameters). It would be of interest to investigate whether this holds true when including other parameters or models. Similar to the time-varying analysis illustrated here for the GGM, future studies could also conduct analyses on only the ascending phase of other models, for example the GLM, as we looked at the entire incidence curve. Further, infectious disease outbreak data, as time series, are naturally correlated, but the fitting performed in our analyses were based on marginal distributions in each time period, assuming independence. Future studies could include a covariance structure while fitting to outbreak data, which would require specification of the correlation structure specific to the disease application.

Overall, the results demonstrate two simple estimation methods that work well and nearly equivalently in the presence of little to no overdispersion, but may have significant uncertainty as the level of overdispersion increases, depending on the amount of available data. It is also shown that more data is needed to provide precise confidence intervals in the presence of increasing levels of overdispersion, which implies that the utilization of more data can resolve potential identifiability issues when high levels of overdispersion is suspected. For both models shown, LSQ with Poisson-MLE provide little to no difference in results with regards to both parameter accuracy and precision.

Acknowledgments

We thank Dr. Ping Yan (Public Health Agency of Canada) for interesting discussions relating to overdispersion and parameter estimation.

References

1. McCullagh P, Nelder JA: *Generalized linear models.* London ; New York : Chapman and Hall, 1989.

2. Williams R: **Heteroskedasticity.** 2015.
3. Dean C, Lundy E: **Wiley StatsRef: Statistics Reference Online.** In *Overdispersion*; 2014.
4. Roosa K, Chowell G: **Assessing parameter identifiability in compartmental dynamic models using a computational approach: Application to infectious disease transmission models.** *Revisions submitted.*; 2018.
5. Efron B, Tibshirani R: *An introduction to the bootstrap.* New York : Chapman & Hall, c1993.; 1993.
6. Chowell G: **Fitting dynamic models to epidemic outbreaks with quantified uncertainty: A primer for parameter uncertainty, identifiability, and forecasts.** *Infectious Disease Modelling* 2017, **2:**379-398.
7. Anderson RM, May RM: *Infectious Diseases of Humans: Dynamics and Control.* 1991.
8. Diekmann O, Heesterbeek JA, Metz JA: **On the definition and the computation of the basic reproduction ratio R0 in models for infectious diseases in heterogeneous populations.** *Journal Of Mathematical Biology* 1990, **28:**365-382.
9. Chowell G, Viboud C, Hyman JM, Simonsen L: **The Western Africa Ebola virus disease epidemic exhibits both global exponential and local polynomial growth rates.** 2014.
10. Viboud C, Simonsen L, Chowell G: **A generalized-growth model to characterize the early ascending phase of infectious disease outbreaks.** *Epidemics* 2016, **15:**27-37.
11. Shanafelt DW, Jones G, Lima M, Perrings C, Chowell G: **Forecasting the 2001 Foot-and-Mouth Disease Epidemic in the UK.** *Ecohealth* 2017.
12. Chowell G, Hincapie-Palacio D, Ospina J, Pell B, Tariq A, Dahal S, Moghadas S, Smirnova A, Simonsen L, Viboud C: **Using Phenomenological Models to Characterize Transmissibility and Forecast Patterns and Final Burden of Zika Epidemics.** Public Library of Science, 2016-05-31.; 2016.

13. Chowell G, Nishiura H: **Comparative estimation of the reproduction number for pandemic influenza from daily case notification data.** *JOURNAL OF THE ROYAL SOCIETY INTERFACE* 2007, **4:**155-166.

14. Dinh L, Chowell G, Rothenberg R: **Growth scaling for the early dynamics of HIV/AIDS epidemics in Brazil and the influence of socio-demographic factors.** *JOURNAL OF THEORETICAL BIOLOGY* 2018, **442:**79-86.

15. Pell B, Kuang Y, Viboud C, Chowell G: **Using phenomenological models for forecasting the 2015 Ebola challenge.** *Epidemics* 2018, **22:**62-70.

16. Ganyani T, Roosa K, Faes C, Hens N, Chowell G: **Assessing the relationship between epidemic growth scaling and epidemic size: The 2014-16 Ebola epidemic in West Africa.** *Epidemiology and Infection* 2018.

17. Dean C, Lundy E: **Overdispersion.** In *Wiley StatsRef: Statistics Reference Online*; 2014.

18. Lewis JM, Lakshmivarahan S, Dhall S: *Dynamic data assimilation: a least squares approach.* Cambridge University Press; 2006.

19. Myung IJ: **Tutorial on maximum likelihood estimation.** vol. 47. pp. 90-100: Journal of Mathematical Pyschology; 2003:90-100.

20. Kashin K: **Statistical Inference: Maximum Likelihood Estimation.** 2014.

Chapter 4. Short-term forecasts of the COVID-19 epidemic in Guangdong and Zhejiang, China: February 13 – 23, 2020

Roosa, K., Lee, Y., Luo, R., Kirpich, A., Rothenberg, R., Hyman, J.M., Yan, P., & Chowell, G. (2020) Short-term forecasts of the COVID-19 epidemic in Guangdong and Zhejiang, China: February 13 – 23, 2020. *Journal of Clinical Medicine.* https://doi.org/10.3390/jcm9020596

1. Introduction

The ongoing epidemic of a novel coronavirus (SARS-CoV-2) began in Hubei Province, China in December 2019 and continues to cause infections in multiple countries, threatening to become a pandemic. However, the bulk of the associated morbidity and mortality is still concentrated within the province of Hubei, China. As of February 13, 2020, a total of 59,907 cumulative cases including 1,368 deaths have been reported globally, with 48,206 cases reported in Hubei alone [1]. To control the epidemic, the Chinese government has enacted a range of social distancing strategies, such as city-wide lockdowns, screening measures at train stations and airports, active case finding, and isolation of suspected cases. However, the numbers of cases and deaths continue to accumulate every day, although transmission appears to be slowing down as a result of strict lockdowns as well as isolation and quarantine measures [1-3].

The epidemiological features of the respiratory illness COVID-19 are still unclear, and changes in reporting of cases and deaths have further complicated analysis of the epidemic, particularly in the epicenter. For instance, the definition of "confirmed" cases has fluctuated over time, and, as of February 12, 2020, has expanded to include clinically suspected cases that have not been laboratory tested. Therefore, the sudden increase in cases observed on February 13th, specifically in Hubei, is attributed to the inclusion of many historical cases that appear to have a sizable gap between onset and reporting. This has obscured the true underlying epidemic trajectory and complicates the inference of epidemiological parameters, such as R_0, and the calibration of mechanistic transmission models.

Phenomenological growth models that capture the empirical patterns of past epidemics can be used to investigate the trajectory of epidemics in real time and are especially useful when the amount of epidemiological data is limited [4-7]. Real-time short-term forecasts generated from such models can be useful to allocate the resources needed to bring the epidemic under control. In this paper, we employ dynamic models to generate 5-day and 10-day ahead forecasts of the cumulative reported cases in the provinces of Guangdong and Zhejiang, China.

2. Methods

2.1. Data

Cumulative case counts by reporting date are reported by the National Health Commission of China and include 34 provinces, including, municipalities, autonomous regions, and special administrative regions [1]. Data were collected daily at 12 pm (GMT-5) from the initial date of reporting, January 22, 2020, to February 13, 2020. Here, we focus on forecasting the trajectory of the epidemic in the provinces of Guangdong and Zhejiang, which have exhibited high burden of COVID-19. We do not forecast the epidemic in Hubei, as the epidemic curve for this province has been distorted as a result of a jump in cases stemming from a change in case reporting on February 13th, 2020.

2.2. Models

We use three phenomenological models that have been previously applied to various infectious disease outbreaks, including other respiratory illnesses such as SARS and pandemic influenza [8, 9], as well as to this current outbreak [10]. The generalized logistic growth model (GLM) and the Richards model extend the simple logistic growth model with an additional scaling parameter [7, 9, 11]. We also apply a sub-epidemic model, which accommodates complex epidemic trajectories, such as multiple peaks and sustained or damped oscillations by assembling the contribution of inferred sub-epidemics [8]. A detailed description of each model and its corresponding parameters is included in Appendix 2.

2.3. Short-term forecasts

We calibrate each of the models to the daily case counts reported for Guangdong and Zhejiang provinces. While we fit to the “incidence” curve, we present results as cumulative case counts.

Reported data are available beginning January 22, 2020, so the calibration period includes daily data from January 22 – February 13, 2020. We estimate the best-fit solution for each model using nonlinear least squares fitting, a process that yields the set of model parameters that minimizes the sum of squared errors between the model and the data. The initial condition is set to the first data point.

We use a parametric bootstrap approach to generate uncertainty bounds around the best-fit solution assuming a Poisson error structure; detailed descriptions of this method are provided in previous works [7, 12]. We refit the models to each of the M = 200 datasets generated by the bootstrap approach, resulting in M best-fit parameter sets that are used to construct the 95% confidence intervals for each parameter. Further, each model solution is used to generate m = 30 additional simulations extended through a 10-day forecasting period. We construct the 95% prediction intervals for forecasts with these 6,000 (M × m) curves.

3. Results

We present results for 5- and 10-day forecasts generated on February 13, 2020 for the provinces of Guangdong and Zhejiang, China. Figures 1 and 2 contain the estimated ranges of cumulative case counts from 5- and 10-day forecasts for Guangdong and Zhejiang, respectively. 10-day ahead forecasts from each model with the calibration data are shown in Figures 3 – 5.

3.1. Guangdong

Our 5-day average forecasts for Guangdong are nearly equivalent across the three models, ranging from 1,290 – 1,304 cumulative reported cases (Figure 1). As of February 13, 2020, Guangdong has a total of 1,241 reported cases [1], so forecasts predict an additional 49 – 63 cases in the next 5 days. Upper bounds (UB) of 95% prediction intervals for both the GLM and Richards model suggest that up to 1,392 cases could accumulate, while the sub-epidemic prediction intervals are substantially wider and include up to 1,699 cases; this translates to an additional 151 – 458 additional cases by February 18, 2020.

10-day forecasts suggest very little increase from the 5-day forecasts, especially for those predicted by the GLM and Richards model (Figure 1). Average 10-day forecasts predict between 1,306 – 1,322 cumulative cases with upper bounds ranging from 1,410 – 1,748 cases. This suggests an additional 65 – 81 cases (UB: 169 – 507) by February 23, 2020.

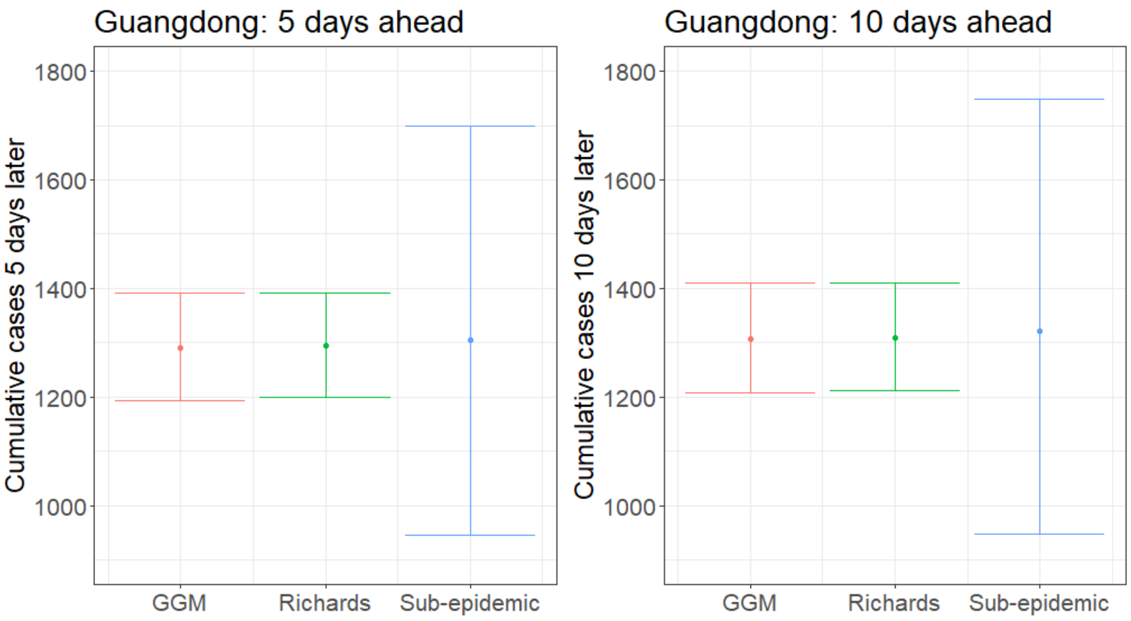


Figure 1. Forecasting results of 5- and 10-days ahead estimates of cumulative reported case counts for Guangdong, China generated on February 13, 2020. The mean estimate for each model is represented by the dots, while the 95% prediction interval is represented by the lines.

3.2. Zhejiang

Average 5-day forecasts from the GLM and Richards model are nearly equivalent for Zhejiang (1,181 and 1,186, respectively), while the sub-epidemic model predicts an average of 1,405 cumulative cases (Figure 2). The sub-epidemic model also has significantly higher upper bounds, suggesting the possibility of up to 1,853 cases, while the GLM and Richards only predict up to 1,276 and 1,279 respectively. As of February 13, 2020, Zhejiang has a total cumulative reported case count of 1,145 [1]; therefore, the models are predicting an additional 36 – 260 cases in the next five days (UB: 131 – 708).

Our 10-day forecasts from the GLM and Richards model show little increase in cases from 5 to 10 days ahead; however, the sub-epidemic model forecasts increase significantly in this time period (Figure 2). 10-day forecasts across the models predict 1,189 – 1,499 cumulative cases, on average, with upper bounds ranging from 1,286 – 2,020 cases. This corresponds to an additional 44 – 354 (UB: 141 – 875) cases in Zhejiang by February 23, 2020.

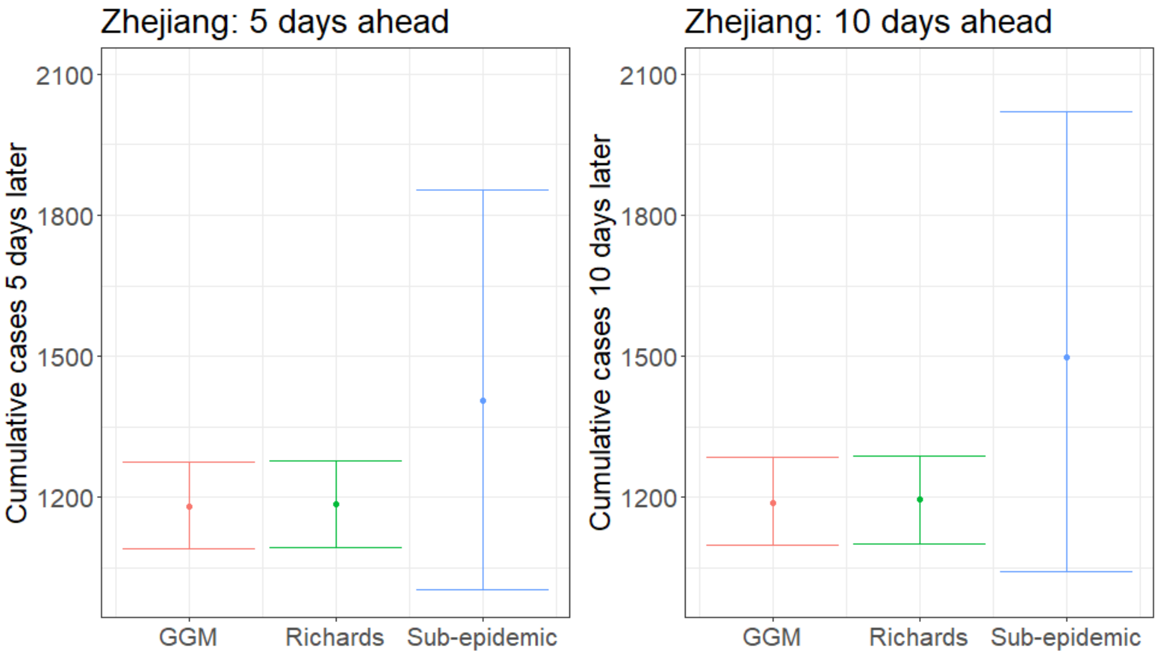


Figure 2. Forecasting results of 5- and 10-days ahead estimates of cumulative reported case counts for Zhejiang, China generated on February 13, 2020. The mean estimate for each model is represented by the dots, while the 95% prediction interval is represented by the lines.

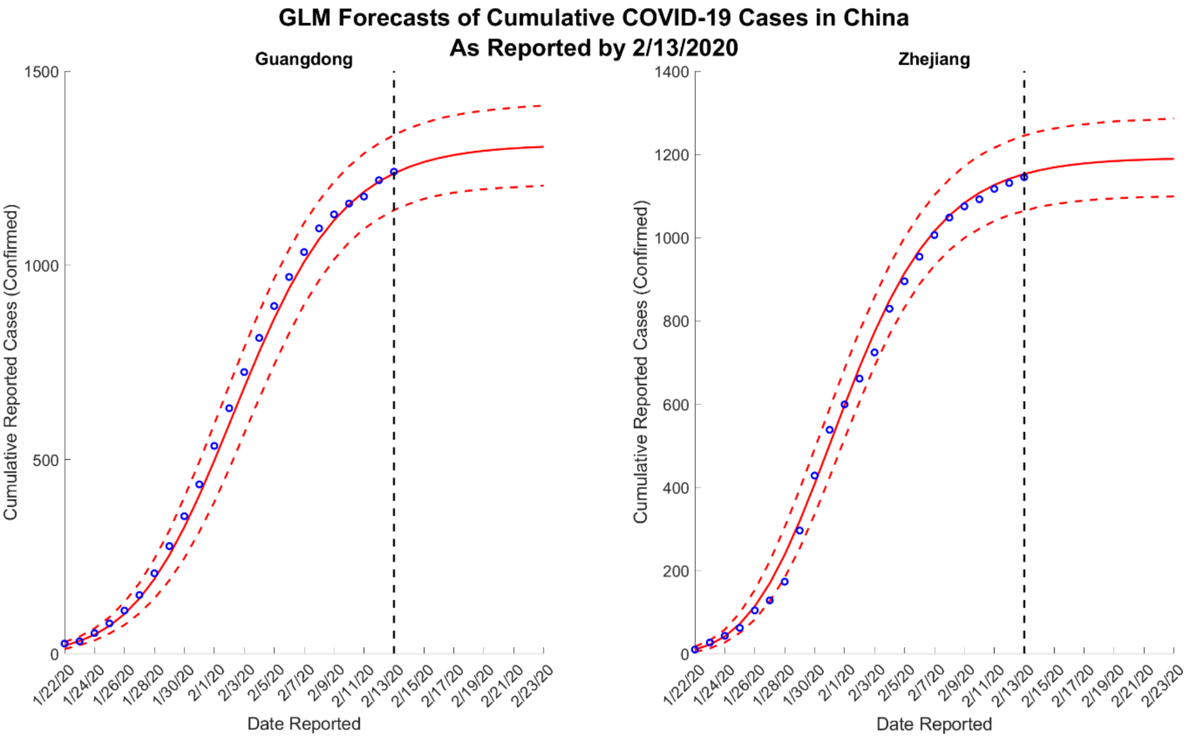


Figure 3. 10-day ahead GLM forecasts of cumulative reported COVID-19 cases in Guangdong and Zhejiang, China – generated on February 13, 2020.

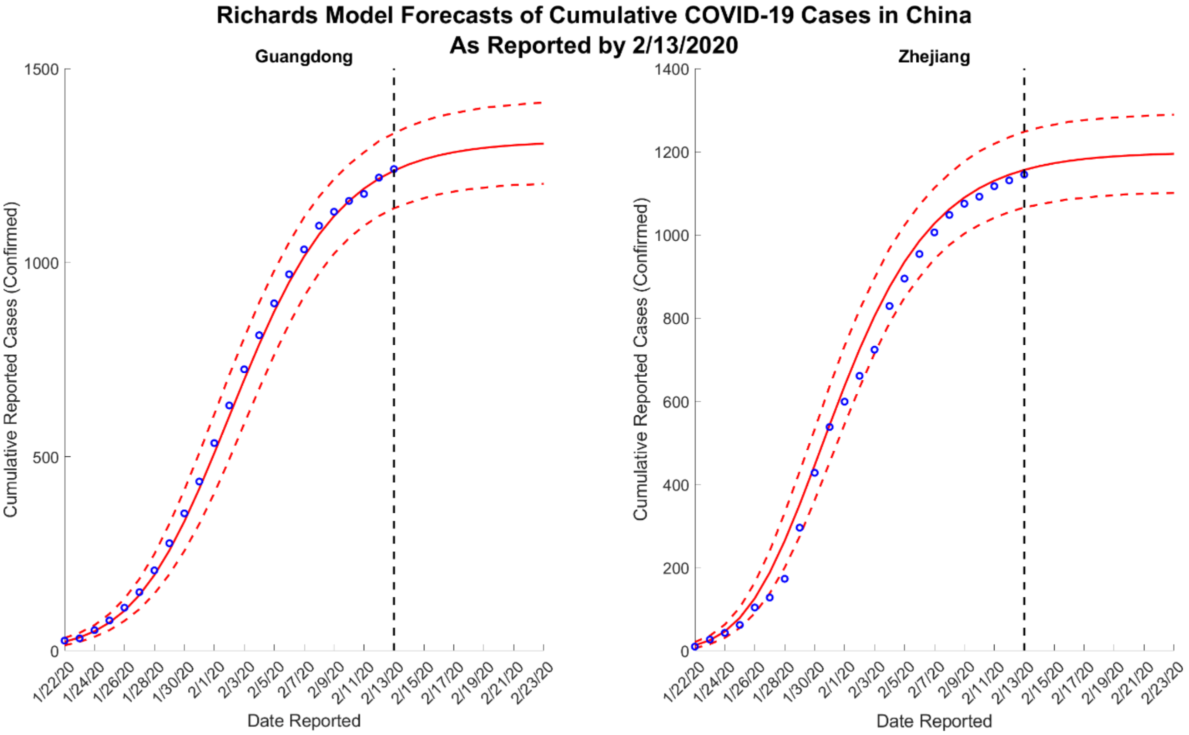


Figure 4. 10-day ahead Richards model forecasts of cumulative reported COVID-19 cases in Guangdong and Zhejiang, China – generated on February 13, 2020.

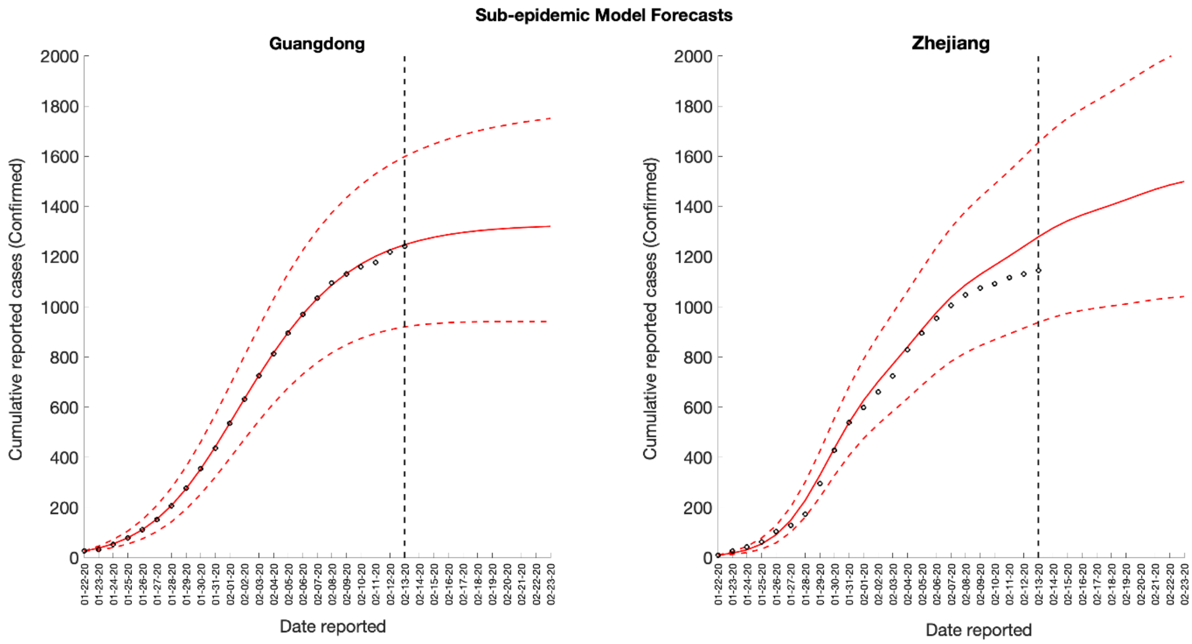


Figure 5. 10-day ahead sub-epidemic model forecasts of cumulative reported COVID-19 cases in Guangdong and Zhejiang, China – generated on February 13, 2020.

4. Discussion

In this report, we present timely short-term forecasts for reported cases of COVID-19 in Guangdong and Zhejiang, China. Based on data reported up to February 13, 2020, the models predict an additional 65 – 81 cases (UB: 169 – 507) in Guangdong and an additional 44 – 354 (UB: 141 – 875) cases in Zhejiang by February 23, 2020. Overall, our forecasts suggest that the epidemic in these two provinces continue to slow down.

Across all forecasts, the GLM and Richards model provide comparable mean estimates and prediction intervals, while the sub-epidemic model forecasts exhibit significantly greater uncertainty (Figures 1 & 2). While the mean estimates for Guangdong are nearly equivalent across all three models, the mean estimates generated by the sub-epidemic model are significantly higher for Zhejiang. Both the GLM and Richards models are predicting that the provinces are nearing the end of the epidemic (Figures 3 & 4). However, forecasts from the sub-epidemic model, which accommodates more complex trajectories, suggest a longer epidemic wave, particularly in Zhejiang (Figure 5).

While we do not know the true underlying epidemic trajectory, it is reasonable to assume the sub-epidemic forecasts better capture the uncertainty for the next 10 days. The fluctuating case definition, particularly, may partially explain the slowing down observed in the data that result in the GLM and Richards model predicting extinction. The kink in the Zhejiang data suggests a case definition change around February 6, 2020, which would partially explain a decrease in the new daily cases reported. The slowing in cases after February 6th is apparent in both Guangdong and Zhejiang; however, this pattern needs to be interpreted with caution. It is not entirely clear whether this is a true decline in transmission or an artificial decline due to the changing case definition. Therefore, the sub-epidemic model forecasts likely better capture both possibilities. Additionally, on February 14, 2020, China officially reported 1,716 cases among healthcare workers that had not been previously identified. The greater potential for transmission by healthcare workers has not been taken into account in this analysis.

In conclusion, while our models predict the outbreaks in Guangdong and Zhejiang have nearly reached extinction, our forecasts need to be interpreted with caution given the unstable case definition and reporting patterns. Thus, we point readers to the sub-epidemic model predictions specifically, which suggest that another wave of cases may occur in the coming days. If the observed decline in case incidence is true, the predictions likely reflect the impact of the social distancing measures implemented by the Chinese government; however, in the best-case scenario, current data suggest that transmission in both provinces is slowing down.

References

1. Chinese National Health Commission. *Reported Cases of 2019-nCoV*. 02/02/2020 - 02/13/2020]; Available from: https://ncov.dxy.cn/ncovh5/view/pneumonia?from=groupmessage&isappinstalled=0.
2. World Health Organization, *Novel Coronavirus (2019-nCoV) Situation Reports*. 2020: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports.
3. Li, Q., et al., *Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia.* The New England Journal Of Medicine, 2020.
4. Pell, B., et al., *Using phenomenological models for forecasting the 2015 Ebola challenge.* Epidemics, 2018. **22**: p. 62-70.

5. Chowell, G., et al., *Using Phenomenological Models to Characterize Transmissibility and Forecast Patterns and Final Burden of Zika Epidemics*. 2016, Public Library of Science, 2016-05-31.

6. Roosa, K., et al., *Multi-model forecasts of the ongoing Ebola epidemic in the Democratic Republic of Congo, March – October 2019*. 2020: Submitted to Royal Society Interface.

7. Chowell, G., *Fitting dynamic models to epidemic outbreaks with quantified uncertainty: A primer for parameter uncertainty, identifiability, and forecasts.* Infectious Disease Modelling, 2017. **2**: p. 379-398.

8. Chowell, G., A. Tariq, and J.M. Hyman, *A novel sub-epidemic modeling framework for short-term forecasting epidemic waves.* BMC Medicine, 2019. **17**(1): p. 164-164.

9. Wang, X.-S., J. Wu, and Y. Yang, *Richards model revisited: Validation by and application to infection dynamics.* Journal of Theoretical Biology, 2012. **313**: p. 12-19.

10. Roosa, K., et al., *Real-time forecasts of the 2019-nCoV epidemic in China from February 5th to February 24th, 2020.* Infectious Disease Modeling, 2020.

11. Richards, F., *A flexible growth function for empirical use.* Journal of Experimental Botany, 1959. **10**(2): p. 290-301.

12. Roosa, K. and G. Chowell, *Assessing parameter identifiability in compartmental dynamic models using a computational approach: Application to infectious disease transmission models.* Theoretical Biology and Medical Modelling, 2019. **16**(1).

13. Viboud, C., L. Simonsen, and G. Chowell, *A generalized-growth model to characterize the early ascending phase of infectious disease outbreaks.* Epidemics, 2016. **15**: p. 27-37.

14. Chowell, G., et al., *Real-time forecasting of epidemic trajectories using computational dynamic ensembles.* Epidemics, 2020. **30**.

15. Roosa, K., R. Luo, and G. Chowell, *Comparative assessment of parameter estimation methods in the presence of overdispersion: a simulation study.* Mathematical Biosciences and Engineering, 2019. **16**(5): p. 4299-4313.

16. Ganyani, T., et al., *Assessing the relationship between epidemic growth scaling and epidemic size: The 2014-16 Ebola epidemic in West Africa.* Epidemiology and Infection, 2018.

Chapter 5. Dissertation summary

Overall, the three studies presented in Chapters 2 – 4 provide a thorough guide for assessing and utilizing mathematical models for describing infectious disease outbreak trends. In the first study, we describe the process for analyzing identifiability of parameters of interest in disease transmission models. We specifically focus on mechanistic models that have been previously applied to infectious disease outbreak scenarios. We use a simple computational approach to assess which parameters can be jointly estimated from the model. We fit the models to data simulated directly from the model, meaning, if the model is structurally identifiable, the estimation process should be able to recover the parameters.

For modeling studies, we recommend conducting comprehensive parameter identifiability analyses based on simulated data prior to attempting to fit the model to real outbreak data. These analyses help guide the set of parameters in the model that can be jointly estimated, as identifiability issues may not arise until any given number of parameters are being simultaneously estimated. If the analysis indicates non-identifiability of certain parameters, it may be necessary to include sensitivity analyses of these estimated parameters. The examples in Chapter 2 provide a guide for conducting these analyses and highlight the importance of assessing identifiability before calibrating a model to real outbreak data.

In the second study, we expand this idea to simple phenomenological models, which are purely empirical and do not rely on disease-specific assumptions of epidemiological parameters. While we utilize a Poisson distribution for the error structure in the first study, we explore the idea of overdispersion and how to determine an appropriate error structure in the second study. We evaluate the effects of misspecification of the data’s error structure on bias and uncertainty associated with parameter estimates using simple dynamic models. Specifically, we focus on modeling varying levels of data overdispersion stemming from randomness in the counting process that shapes the time series data, rather than systematic misspecifications in the mean process linked to the dynamic model.

We show that more data is needed to provide precise confidence intervals in the presence of increasing levels of overdispersion, which implies that the utilization of more data can resolve potential identifiability issues when high levels of overdispersion is suspected. For both models shown, nonlinear least squares and Poisson-MLE provide little to no difference in parameter estimation results with regards to both parameter accuracy and precision.

In the third study, we utilize previously validated phenomenological models, including one analyzed in the second study, to generate short-term forecasts of the ongoing COVID-19 pandemic. We show that simple phenomenological models are successful for fitting and forecasting real disease outbreak trends. During infectious disease epidemics, public health authorities rely on modeling results to inform intervention decisions and resource allocation. Therefore, we highlight the importance of interpreting modeling results with caution, particularly given the quality of data during the outbreak. We also highlight the sub-epidemic modeling framework, as it allows for dynamics that suggest another wave of cases may occur; whereas, single peak models cannot predict resurgences.

In summary, the ability to make sound public health decisions regarding an infectious disease outbreak is crucial for the general health and safety of a population. Knowledge of whether a parameter is identifiable from a given model and data is invaluable, as estimates of non-identifiable parameters should not be used to inform public health decisions. Theoretical aspects of mathematical models and parameter estimation methods must be considered before applying to real outbreaks, and the underlying assumptions of the model should be presented clearly. Further, results from modeling studies should be presented with quantified uncertainty and interpreted in terms of the assumptions and limitations of the model, methods, and data used. The methodology presented in this dissertation provides a thorough guide for conducting model-based inferences.

References for Chapter 1

1. Anderson RM, May RM. Infectious Diseases of Humans: Dynamics and Control1991.

2. Diekmann O, Heesterbeek JA, Metz JA. On the definition and the computation of the basic reproduction ratio R0 in models for infectious diseases in heterogeneous populations. Journal Of Mathematical Biology. 1990;28(4):365-82.

3. Chowell G. Fitting dynamic models to epidemic outbreaks with quantified uncertainty: A primer for parameter uncertainty, identifiability, and forecasts. Infectious Disease Modelling. 2017;2:379-98.

4. Kao Y-H, Eisenberg M. Practical unidentifiability of a simple vector-borne model: implications for parameter estimation and intervention assessment. Epidemics. 2018.

5. Roosa K, Chowell G. Assessing parameter identifiability in compartmental dynamic models using a computational approach: Application to infectious disease transmission models. Theoretical Biology and Medical Modelling. 2019;16(1).

6. McCullagh P, Nelder JA. Generalized linear models: London ; New York : Chapman and Hall, 1989.

7. Dean C, Lundy E. Wiley StatsRef: Statistics Reference Online. Overdispersion2014.

8. Roosa K, Luo R, Chowell G. Comparative assessment of parameter estimation methods in the presence of overdispersion: a simulation study. Mathematical Biosciences and Engineering. 2019;16(5):4299-313.

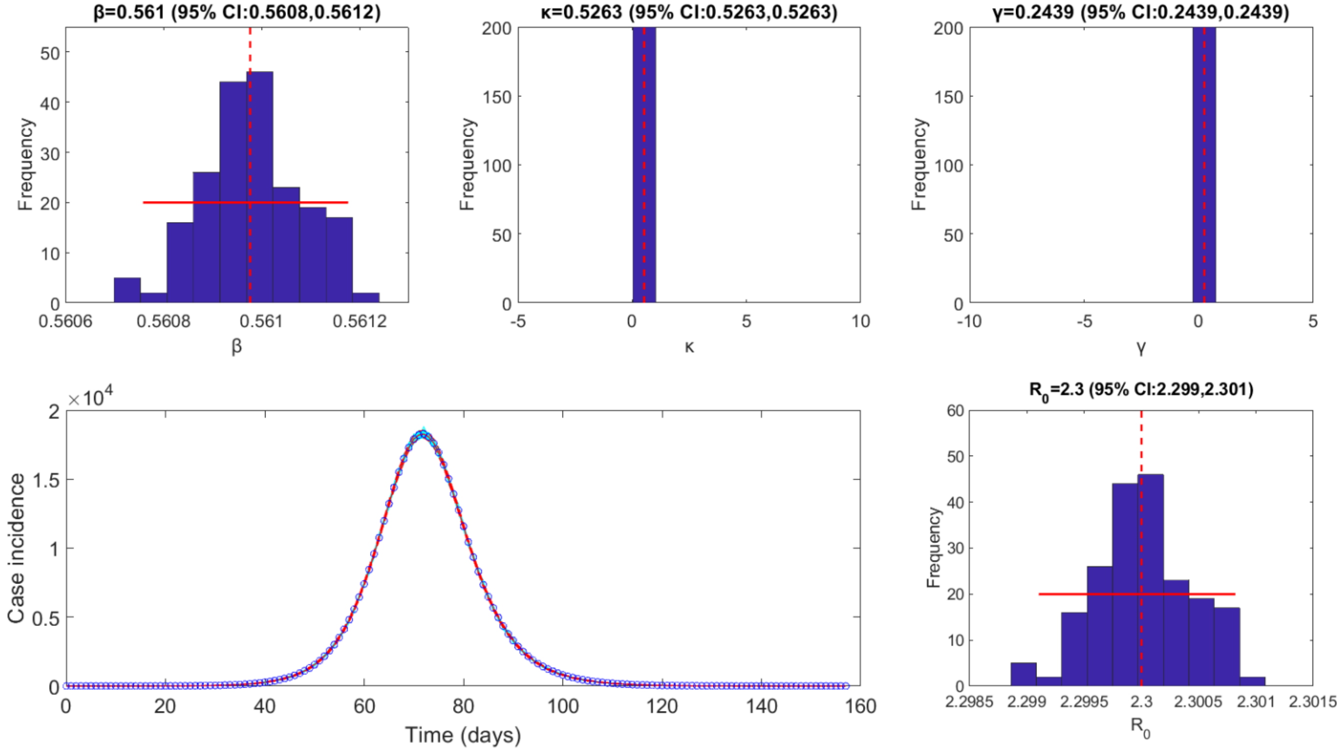
9. Pell B, Kuang Y, Viboud C, Chowell G. Using phenomenological models for forecasting the 2015 Ebola challenge. Epidemics. 2018;22:62-70.

10. Chowell G, Hincapie-Palacio D, Ospina J, Pell B, Tariq A, Dahal S, et al. Using Phenomenological Models to Characterize Transmissibility and Forecast Patterns and Final Burden of Zika Epidemics. Public Library of Science, 2016-05-31.; 2016.

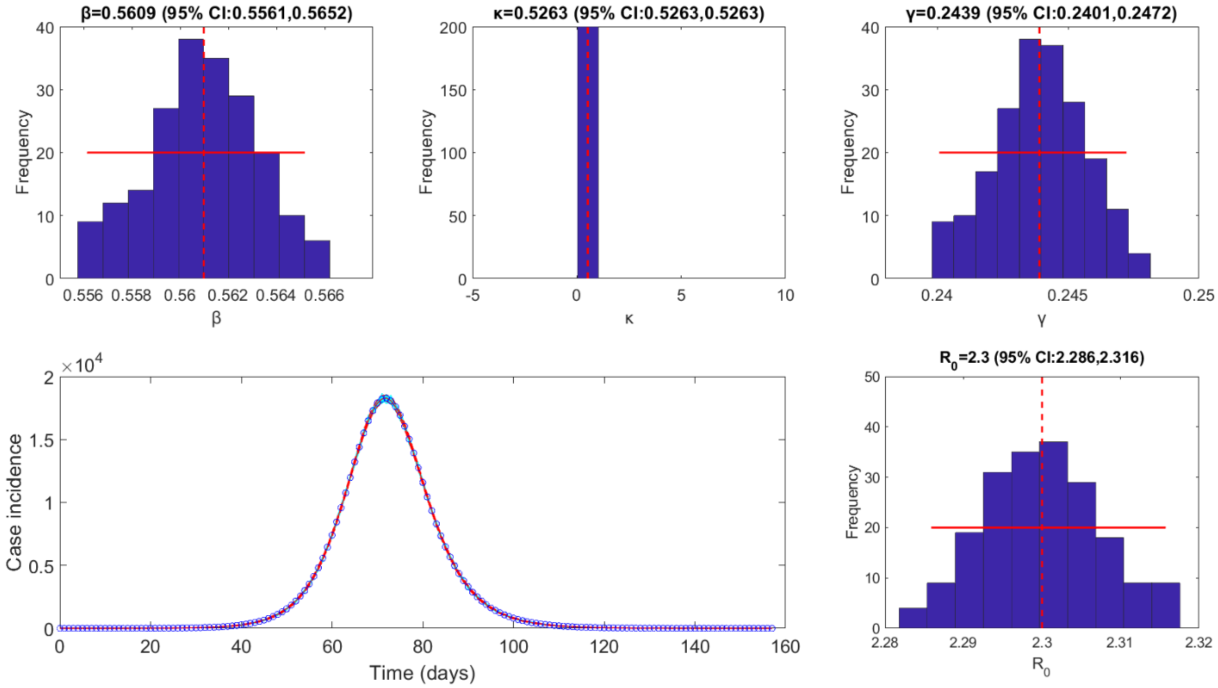
11. Roosa K, Tariq A, Yan P, Hyman J, Chowell G. Multi-model forecasts of the ongoing Ebola epidemic in the Democratic Republic of Congo, March – October 2019. 2020.

12. Roosa K, Lee Y, Luo R, Kirpich A, Rothenberg R, Hyman JM, et al. Short-term forecasts of the COVID-19 epidemic in Guangdong and Zhejiang, China: February 13-23, 2020. Journal of Clinical Medicine. 2020;9(2):596.

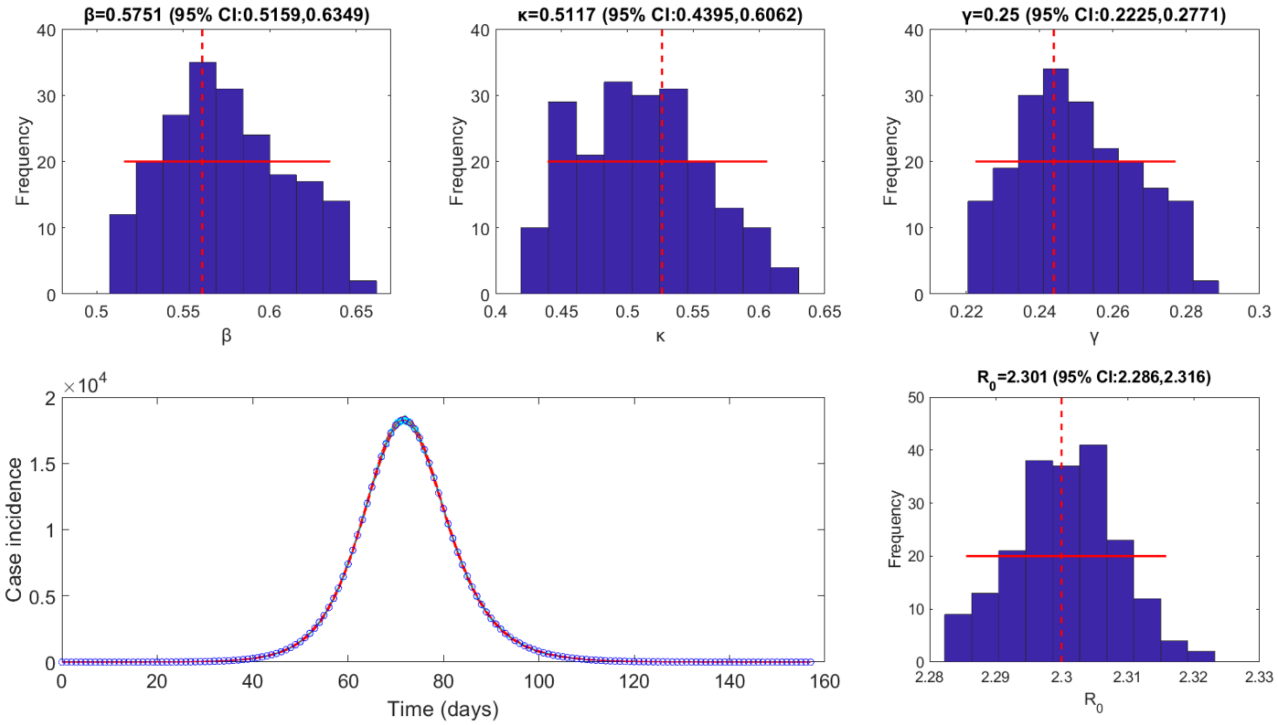
Appendix 1. Supplemental figures for Chapter 2: Assessing parameter identifiability in compartmental dynamic models using a computational approach: application to infectious disease transmission models



Supplemental Figure S1. Model 1 – Θ_1 (estimating β only): The histograms display the empirical distributions of the parameter estimates using 200 bootstrap realizations, where the solid red horizontal line represents the 95% confidence interval for parameter estimates, and the dashed red vertical line indicates the true parameter value. Note, κ and γ are set to their true values in the data. The bottom left graph shows the data from the model (blue circles), and 200 realizations of the epidemic curve assuming a Poisson error structure (light blue lines). The solid red line corresponds to the best-fit of the model to the data, and the dashed red lines correspond to the 95% confidence bands around the best fit.



Supplemental Figure S2. Model 1 – Θ_2 (estimating β and γ): The histograms display the empirical distributions of the parameter estimates using 200 bootstrap realizations, where the solid red horizontal line represents the 95% confidence interval for parameter estimates, and the dashed red vertical line indicates the true parameter value. Note, κ is set to the true value from the data. The bottom left graph shows the data from the model (blue circles), and 200 realizations of the epidemic curve assuming a Poisson error structure (light blue lines). The solid red line corresponds to the best-fit of the model to the data, and the dashed red lines correspond to the 95% confidence bands around the best fit.



Supplemental Figure S3. Model 1 – Θ_3 (estimating β, κ, and γ): The histograms display the empirical distributions of the parameter estimates using 200 bootstrap realizations, where the solid red horizontal line represents the 95% confidence interval for parameter estimates, and the dashed red vertical line indicates the true parameter value. The bottom left graph shows the data from the model (blue circles), and 200 realizations of the epidemic curve assuming a Poisson error structure (light blue lines). The solid red line corresponds to the best-fit of the model to the data, and the dashed red lines correspond to the 95% confidence bands around the

Appendix 2. Supplemental material for Chapter 4: Short-term forecasts of the COVID-19 epidemic in Guangdong and Zhejiang, China: February 13 – 23, 2020

Generalized logistic growth model

The generalized logistic growth model (GLM) extends the simple logistic growth model with a scaling of growth parameter *p* that accommodates sub-exponential growth patterns [1-4]. The GLM is defined by the differential equation:

$$C'(t) = rC(t)^p(1 - \frac{C(t)}{K})$$

where *C(t)* is the cumulative cases at time *t*, *r* is the early growth rate, *p* is the scaling of growth parameter, and *K* is the carrying capacity or final epidemic size. Values of *p* = 1 correspond with exponential growth, *p* = 0 represents constant growth, and 0 < *p* < 1 defines sub-exponential growth.

Richards model

The Richards model also extends the simple logistic growth model through a scaling parameter, *a* that measures the deviation from the symmetric simple logistic curve [5, 6, 7]. The Richards model is defined by the differential equation:

$$\boldsymbol{C'(t) = rC(t)\left(1 - (\frac{C(t)}{K})^a\right)}$$

where *C(t)* represents the cumulative case count at time *t*, *r* is the growth rate, *K* is the final epidemic size, and *a* is a scaling parameter.

Sub-epidemic model

While the GLM and Richards model only accommodate s-shaped dynamics, the sub-epidemic wave model supports complex epidemic trajectories. For this approach, we assume that the observed curve is the aggregate of multiple overlapping sub-epidemics, where each sub-epidemic is modeled using the GLM [8]. An epidemic wave composed of n overlapping sub-epidemics is modeled as follows:

$$\boldsymbol{C_i{}'(t) = rA_{i-1}(t)C_i(t)^p\left(1 - \frac{C_i(t)}{K_i}\right)}$$

where *C*i*(t)* is the cumulative number of infections in sub-epidemic *i* (*i* = 1, …, *n*), *K*i is the size of the *i*th sub-epidemic, and the growth rate *r* and scaling parameter *p* are the same across sub-epidemics [8]. Further, when *n* = 1, the model returns to the single-equation GLM as presented above.

The timing of onset for each consecutive sub-epidemic is modeled with a regular structure, such that the (*i*+1)th sub-epidemic is triggered when the cumulative case count of sub-epidemic *i*, *C*i*(t)*, exceeds the threshold C_{thr}. The (*i*+1)th sub-epidemic begins before the *i*th sub-epidemic reaches extinction. The size of consecutive sub-epidemics (K_i) is modeled such that the size declines exponentially for each subsequent sub-epidemic, where

$$\boldsymbol{K_i = K_0 e^{-q(i-1)}}$$

and K_0 is the size of the first sub-epidemic ($K_1 = K_0$), and *q* is the rate of decline, where *q* = 0 corresponds to no decline. The total final epidemic size is given by:

$$\boldsymbol{K_{tot} = \sum_{i=1}^{n_{tot}} K_0 e^{-q(i-1)} = \frac{K_0(1 - e^{-qn_{tot}})}{1 - e^{-q}}}$$

where n_{tot} is the finite number of overlapping sub-epidemics, calculated as

$$\boldsymbol{n_{tot} = \left\lfloor -\frac{1}{q}\ln\left(\frac{C_{thr}}{K_0}\right) + 1\right\rfloor}$$

References

1. Viboud, C., L. Simonsen, and G. Chowell, A generalized-growth model to characterize the early ascending phase of infectious disease outbreaks. Epidemics, 2016. 15: p. 27-37.

2. Chowell, G., et al., Real-time forecasting of epidemic trajectories using computational dynamic ensembles. Epidemics, 2020. 30.

3. Roosa, K., R. Luo, and G. Chowell, Comparative assessment of parameter estimation methods in the presence of overdispersion: a simulation study. Mathematical Biosciences and Engineering, 2019. 16(5): p. 4299-4313.

4. Ganyani, T., et al., Assessing the relationship between epidemic growth scaling and epidemic size: The 2014-16 Ebola epidemic in West Africa. Epidemiology and Infection, 2018.

5. Chowell, G., Fitting dynamic models to epidemic outbreaks with quantified uncertainty: A primer for parameter uncertainty, identifiability, and forecasts. Infectious Disease Modelling, 2017. 2: p. 379-398.

6. Wang, X.-S., J. Wu, and Y. Yang, Richards model revisited: Validation by and application to infection dynamics. Journal of Theoretical Biology, 2012. 313: p. 12-19.

7. Richards, F., A flexible growth function for empirical use. Journal of Experimental Botany, 1959. 10(2): p. 290-301.

8. Chowell, G., A. Tariq, and J.M. Hyman, A novel sub-epidemic modeling framework for short-term forecasting epidemic waves. BMC Medicine, 2019. 17(1): p. 164-164.