

RESEARCH

Open Access



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

Kimberlyn Roosa^{1*}  and Gerardo Chowell^{1,2}

Abstract

Background: Mathematical modeling is now frequently used in outbreak investigations to understand underlying mechanisms of infectious disease dynamics, assess patterns in epidemiological data, and forecast the trajectory of epidemics. However, the successful application of mathematical models to guide public health interventions lies in the ability to reliably estimate model parameters and their corresponding uncertainty. Here, we present and illustrate a simple computational method for assessing parameter identifiability in compartmental epidemic models.

Methods: We describe a parametric bootstrap approach to generate simulated data from dynamical systems to quantify parameter uncertainty and identifiability. We calculate confidence intervals and mean squared error of estimated parameter distributions to assess parameter 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.

Results: Overall, parameter identifiability issues are more likely to arise with more complex models (based on number of equations/states and parameters). As the number of parameters being jointly estimated increases, the uncertainty surrounding estimated parameters tends to increase, on average, as well. We found that, in most cases, R_0 is often robust to parameter identifiability issues affecting individual parameters in the model. Despite large confidence intervals and higher mean squared error of other individual model parameters, R_0 can still be estimated with precision and accuracy.

Conclusions: Because public health policies can be influenced by results of mathematical modeling studies, it is important to conduct parameter identifiability analyses prior to fitting the models to available data and to report parameter estimates with quantified uncertainty. The method described is helpful in these regards and enhances the essential toolkit for conducting model-based inferences using compartmental dynamic models.

Keywords: Compartmental models, Parameter identifiability, Uncertainty quantification, Epidemic models, Structural parameter identifiability, Practical parameter identifiability

* Correspondence: kroosa1@student.gsu.edu

¹Department of Population Health Sciences, School of Public Health, Georgia State University, Atlanta, GA, USA

Full list of author information is available at the end of the article



Background

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.

Methods

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:

$$\begin{aligned}\dot{x}_1(t) &= f_1(x_1, x_2, \dots, x_h, \Theta) \\ \dot{x}_2(t) &= f_2(x_1, x_2, \dots, x_h, \Theta) \\ &\vdots \\ \dot{x}_h(t) &= f_h(x_1, x_2, \dots, x_h, \Theta)\end{aligned}$$

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

Model 1: Simple SEIR (pandemic influenza)

We analyze a simple compartmental transmission model that consists of 4 parameters and 4 states (Fig. 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.

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: $R_0 = \frac{\beta}{\gamma}$.

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 (Fig. 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 $(1 - \rho)$ 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 .

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

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

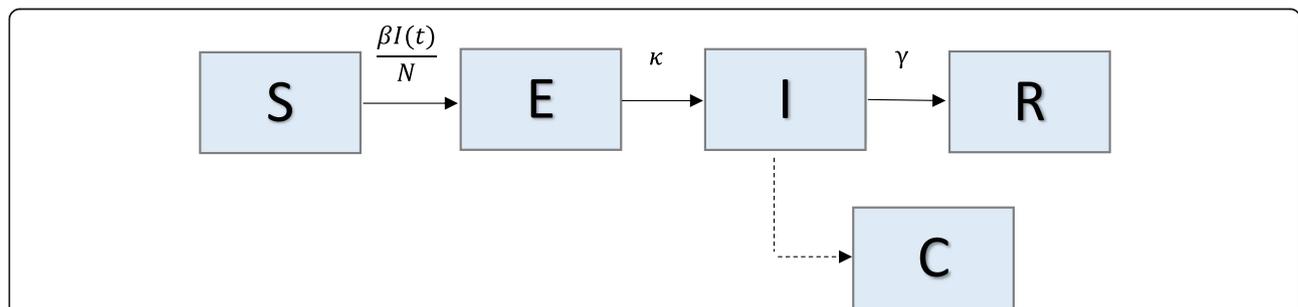


Fig. 1 Model 1: Simple SEIR – Population is divided into 4 classes: susceptible (S), exposed (E), infectious (I), and recovered/removed (R). Class C represents the auxiliary variable $C(t)$ and tracks the cumulative number of infectious individuals from the start of the outbreak. This is presented as 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

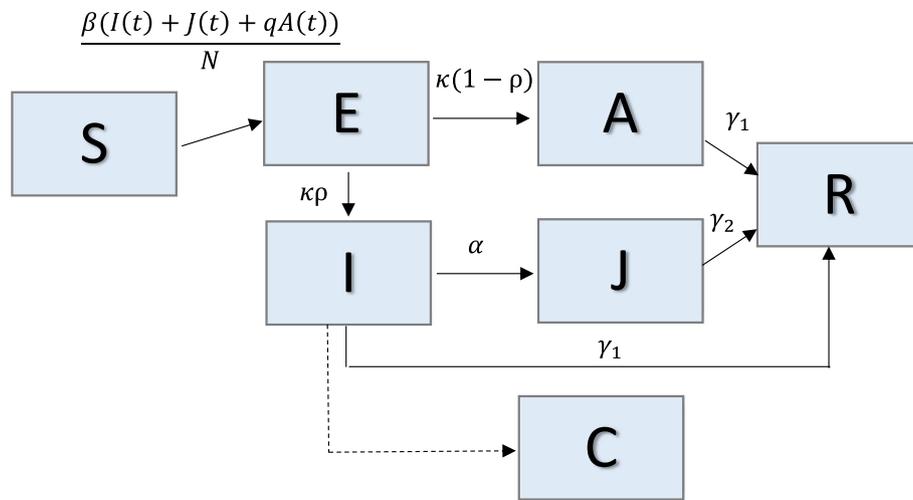


Fig. 2 Model 2: SEIR with asymptomatic and hospitalized/diagnosed and reported – Population is divided into 6 classes: susceptible (S), exposed (E), clinically ill and infectious (I), asymptomatic and partially infectious (A), hospitalized/diagnosed and reported (J), and recovered (R). Class C represents 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

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:

$R_0^A = (\text{fraction of asymptomatic cases}) \times (\text{transmission rate}) \times (\text{relative transmissibility from asymptomatic cases}) \times (\text{mean time in asymptomatic class})$

$R_0^I = (\text{fraction of symptomatic cases}) \times (\text{transmission rate}) \times (\text{mean time in clinically infectious class})$

$R_0^J = (\text{fraction of symptomatic cases that are hospitalized}) \times (\text{transmission rate}) \times (\text{mean time in hospital})$ [24]

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

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

We analyze an Ebola transmission model [25] comprised of 15 parameters and 6 states (Fig. 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 $(1-\theta)$, a proportion, $0 < \delta_1 < 1$, move to the funeral class at rate γ_d , and the rest $(1-\delta_1)$ 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 $\gamma_{dh} = \frac{1}{\gamma_d + \gamma_h}$.

The remaining proportion $(1-\delta_2)$ are recovered/removed at rate $\gamma_{ih} = \frac{1}{\gamma_i + \gamma_h}$. 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 .

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:

$R_0^I = (\text{transmission rate in the community}) \times (\text{mean time in infectious class})$

$R_0^H = (\text{fraction of hospitalized cases}) \times (\text{transmission rate in the hospital}) \times (\text{mean time in hospital class})$

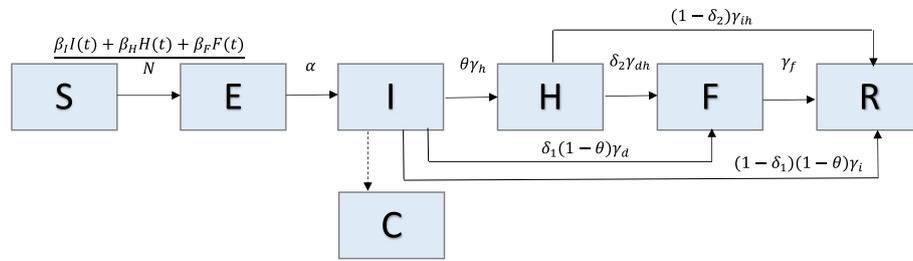


Fig. 3 Model 3: The Legrand et al. Model – Population is divided into 6 classes: susceptible (S), exposed (E), infectious in the community (I), infectious in the hospital (H), infectious after death at funeral (F), or recovered/removed (R). Class C represents the auxiliary variable C(t) and tracks the 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

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

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

where $\Delta = \gamma_h \theta + \gamma_d (1 - \theta) \delta_1 + \gamma_i (1 - \theta) (1 - \delta_1)$ [25].

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 (Fig. 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), asymptotically 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 $S_v + E_v + I_v = N_v$.

A proportion $0 < \theta < 1$ of susceptible humans move to the exposed class at rate $ab(I_v(t)/N_h) + \beta[(\alpha E_h(t) + I_{h1}(t) + \tau I_{h2}(t))/N_h]$ where a is the mosquito biting

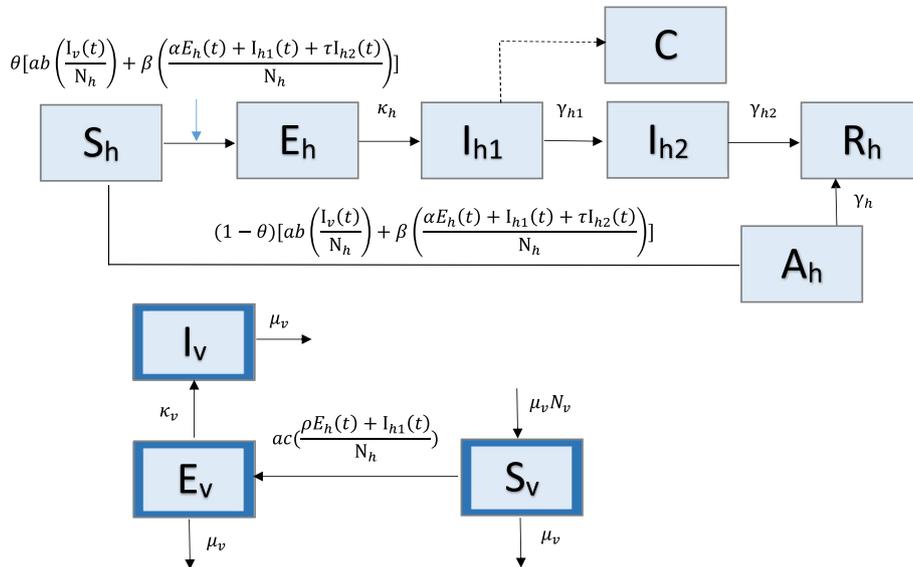


Fig. 4 Model 4: Zika Model with human and mosquito populations – The human population (subscript h) is divided into 5 classes: susceptible (S_h), asymptotically infected (A_h), exposed (E_h), symptomatically infectious (I_{h1}), convalescent (I_{h2}), or recovered (R_h). Class C represents the auxiliary 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

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 $(1 - \theta)$ become asymptotically infected at the same rate, $ab(I_v(t)/N_h) + \beta[(\alpha E_h(t) + I_{h1}(t) + \tau I_{h2}(t))/N_h]$. 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 $ac[(\rho E_h(t) + I_{h1}(t))/N_h]$, 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:

$$\left\{ \begin{array}{l} \dot{S}_h(t) = -ab(I_v(t)/N_h)S_h(t) \\ \quad -\beta[(\alpha E_h(t) + I_{h1}(t) + \tau I_{h2}(t))/N_h]S_h(t) \\ \dot{E}_h(t) = \theta[ab(I_v(t)/N_h)S_h(t) + \beta[(\alpha E_h(t) + I_{h1}(t) \\ \quad + \tau I_{h2}(t))/N_h]S_h(t)] - \kappa_h E_h(t) \\ \dot{I}_{h1}(t) = \kappa_h E_h(t) - \gamma_{h1} I_{h1}(t) \\ \dot{I}_{h2}(t) = \gamma_{h1} I_{h1}(t) - \gamma_{h2} I_{h2}(t) \\ \dot{A}_h(t) = (1-\theta)[ab(I_v(t)/N_h)S_h(t) + \beta[\alpha E_h(t) \\ \quad + I_{h1}(t) + \tau I_{h2}(t))/N_h]S_h(t)] - \gamma_h A_h(t) \\ \dot{R}_h(t) = \gamma_{h2} I_{h2}(t) + \gamma_h A_h(t) \\ \dot{S}_v(t) = \mu_v N_v - ac[(\rho E_h(t) + I_{h1}(t))/N_h] \\ \quad * S_v(t) - \mu_v S_v(t) \\ \dot{E}_v(t) = ac[(\rho E_h(t) + I_{h1}(t))/N_h] * S_v(t) \\ \quad - (\kappa_v + \mu_v) E_v(t) \\ \dot{I}_v(t) = \kappa_v E_v(t) - \mu_v I_v(t) \\ \dot{C}(t) = \kappa_h E_h(t) \end{array} \right.$$

$C(t)$ 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:

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

The overall basic reproductive number, considering both transmission routes, is given by the following eq. [26]:

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

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, 2, 3 and 4.

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 $\hat{\Theta} = (\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_m)$ 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.

Table 1 Parameter descriptions and values for Model 1

Parameters	Description	Value
N	Population size	500,000
β	Transmission rate (per day)	0.56
$1/\kappa$	Mean latent period (days)	1.9
$1/\gamma$	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	500,000
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]

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.

Table 3 Parameter descriptions and values for Model 3

Parameters	Description	Value
N	Population size	200,000
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/ γ_h	Time from symptom onset to hospitalization (days)	5
1/ γ_d	Time from symptom onset to death (days)	9.6
1/ γ_i	Time from symptom onset to the end of infectiousness for survivors (days)	10
δ	Case fatality ratio	0.81
δ_1	$\delta_1 = \frac{\delta \gamma_i}{\delta \gamma_i + (1-\delta) \gamma_d}$	0.80
δ_2	$\delta_2 = \frac{\delta \gamma_h}{\delta \gamma_h + (1-\delta) \gamma_{dh}}$	0.80
1/ γ_{ih}	Infectious period for survivors (days)	5
1/ γ_{dh}	Time from hospitalization to death (days)	4.6
1/ γ_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)	200,000
N_v	Population size (mosquitos)	1,000,000
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
	Proportion of symptomatic infections	0.18
1/ κ_h	Intrinsic incubation period in humans (days)	5
1/ γ_{h1}	Duration of acute phase (days)	5
1/ γ_{h2}	Duration of convalescent phase (days)	20
1/ γ_h	Duration of asymptomatic infection (days)	
1/ μ_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/ κ_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]

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*

$$\Theta_i : \begin{aligned} \Theta_1 &= \{ \hat{\beta}, \kappa, \gamma \} \\ \Theta_2 &= \{ \hat{\beta}, \kappa, \hat{\gamma} \} \\ \Theta_3 &= \{ \hat{\beta}, \hat{\kappa}, \hat{\gamma} \} \end{aligned}$$

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

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

$$\Theta_2 = \left\{ \hat{\beta}, \kappa, \hat{\gamma}_1, \gamma_2, \alpha, \rho, q \right\}$$

$$\Theta_3 = \left\{ \hat{\beta}, \kappa, \hat{\gamma}_1, \gamma_2, \hat{\alpha}, \rho, q \right\}$$

$$\Theta_4 = \left\{ \hat{\beta}, \kappa, \hat{\gamma}_1, \gamma_2, \hat{\alpha}, \hat{\rho}, \hat{q} \right\}$$

$$\Theta_5 = \left\{ \hat{\beta}, \kappa, \hat{\gamma}_1, \gamma_2, \hat{\alpha}, \hat{\rho}, \hat{q} \right\}$$

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

$$\Theta_i : \Theta_1 = \left\{ \hat{\beta}_I, \hat{\beta}_H, \hat{\beta}_F, \alpha, \theta, \gamma_h, \gamma_d, \gamma_i, \delta, \gamma_{ih}, \gamma_{dh}, \gamma_f \right\}$$

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

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

$$\Theta_4 = \left\{ \hat{\beta}_I, \hat{\beta}_H, \hat{\beta}_F, \alpha, \theta, \hat{\gamma}_h, \hat{\gamma}_d, \hat{\gamma}_i, \delta, \gamma_{ih}, \gamma_{dh}, \gamma_f \right\}$$

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

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

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

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

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

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

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

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

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

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

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 $i = 1, 2, \dots, S$.
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.

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: $MSE = \frac{1}{S} \sum_{i=1}^S (\theta - \hat{\theta}_i)^2$ where θ represents the true parameter value (in the simulated data), and $\hat{\theta}_i$ represents the estimated value of the parameter for the i^{th} 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.

Results

Model 1: Simple SEIR

Additional files 1, 2 and 3: illustrate the empirical distributions of the estimated parameters, where Additional file 1: represents the results for $\hat{\Theta}_1$ (β only), Additional file 2: for $\hat{\Theta}_2$ (β and γ), and Additional file 3: for $\hat{\Theta}_3$ (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 Fig. 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 (Fig. 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 .

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.

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 (Figs. 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, $7.15 \cdot 10^{-4}$, 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 $< 10^{-2}$.

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 $> 10^{-2}$ (Fig. 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 (Figs. 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

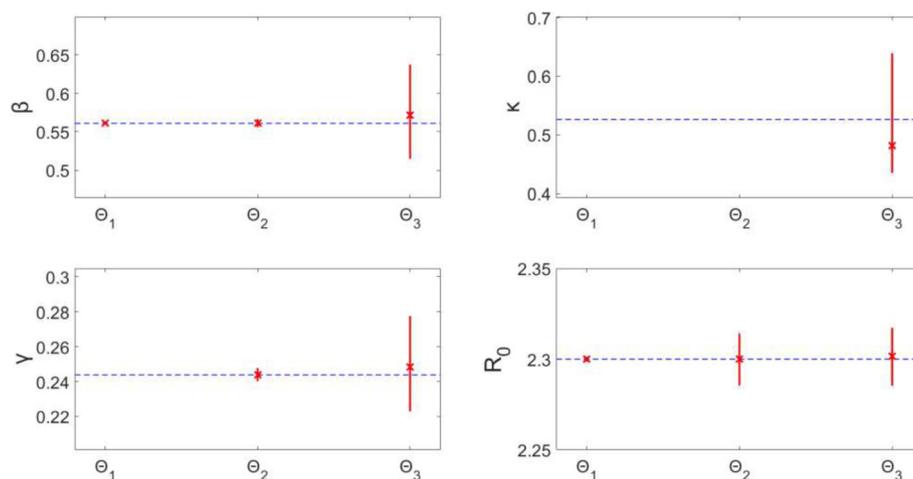
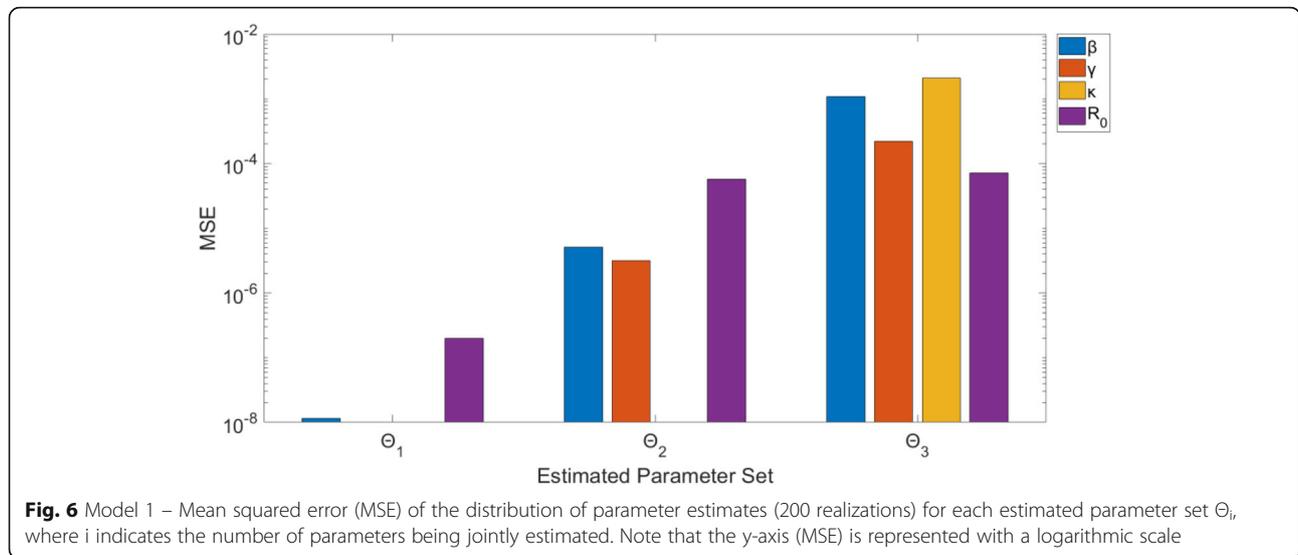


Fig. 5 Model 1–95% confidence intervals (vertical red lines) for the distributions of each estimated parameter obtained from the 200 realizations of the simulated datasets. Mean estimated parameter value is denoted by a red x, and the true parameter value is represented by the blue dashed horizontal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



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.

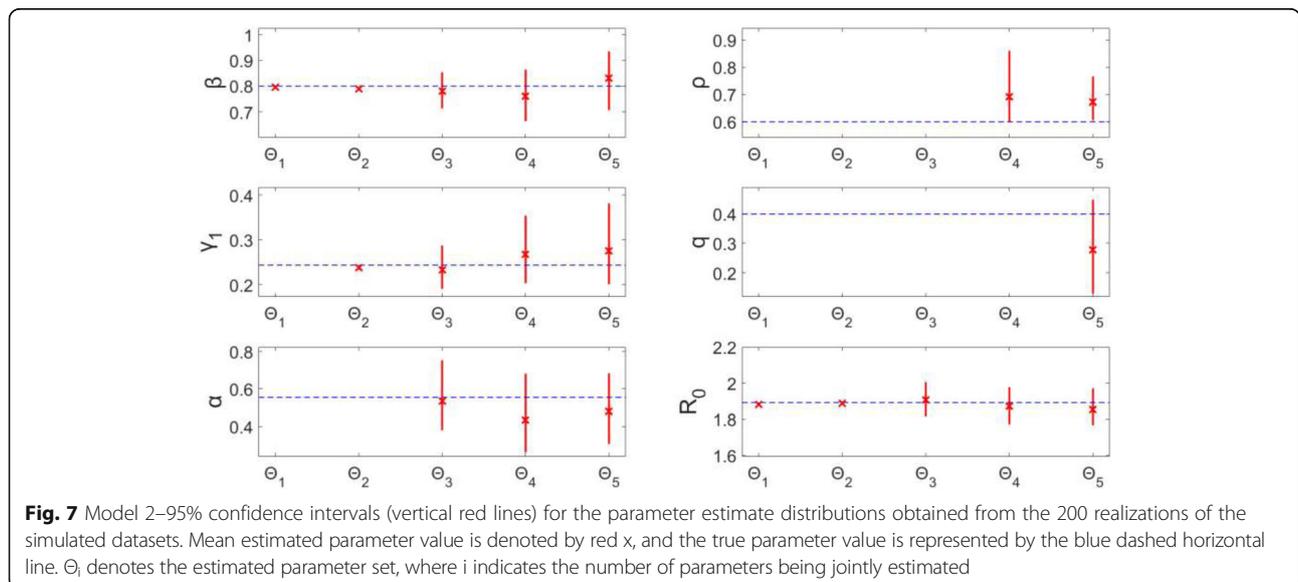
Model 3: The Legrand model (Ebola)

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 (Figs. 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 ($MSE = 2.71 \cdot 10^{-4}$). 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 (Fig. 9).

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 $< 10^{-2}$ (Fig. 10), but the wide range of uncertainty around these parameters is still indicative of non-identifiability (Fig. 9).

Remarkably, R_0 can be precisely estimated with unbiased results for parameter sets $\Theta_1 - \Theta_4$ (Figs. 9 & 10). When



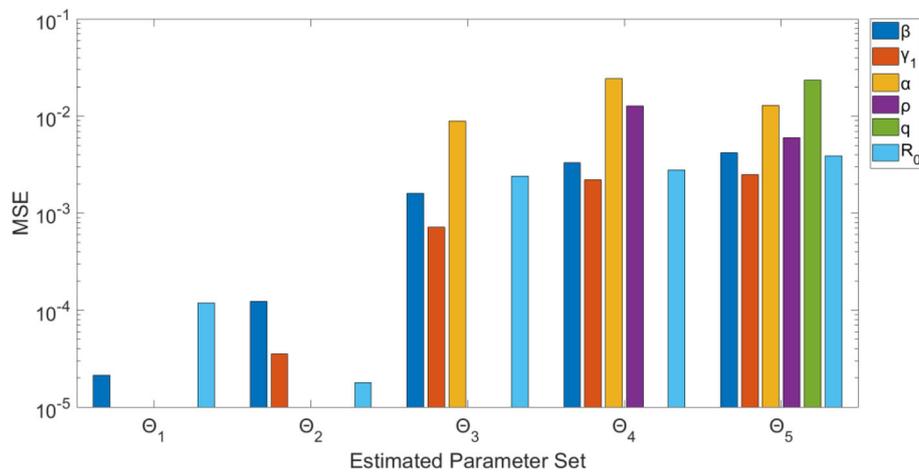


Fig. 8 Model 2 – Mean squared error (MSE) of the distribution of parameter estimates (200 realizations) for each estimated parameter set Θ_i , where i indicates the number of parameters being jointly estimated. Note that the y-axis (MSE) is represented with a logarithmic scale

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.

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 (Figs. 11

& 12). When jointly estimating more than two parameters ($\Theta_i: i > 2$), 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 other parameters ($\Theta_i: i > 2$) 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.

In terms of R_0 , we can see that this composite parameter of interest is identifiable for all Θ_i (Figs. 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

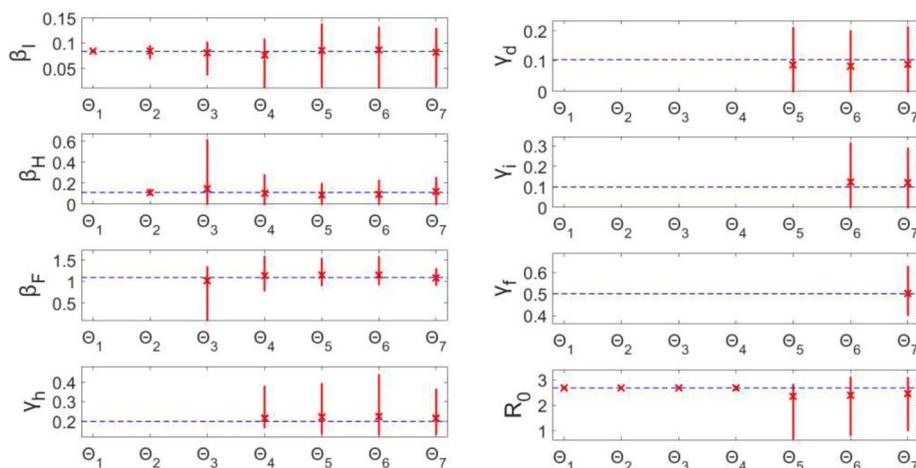
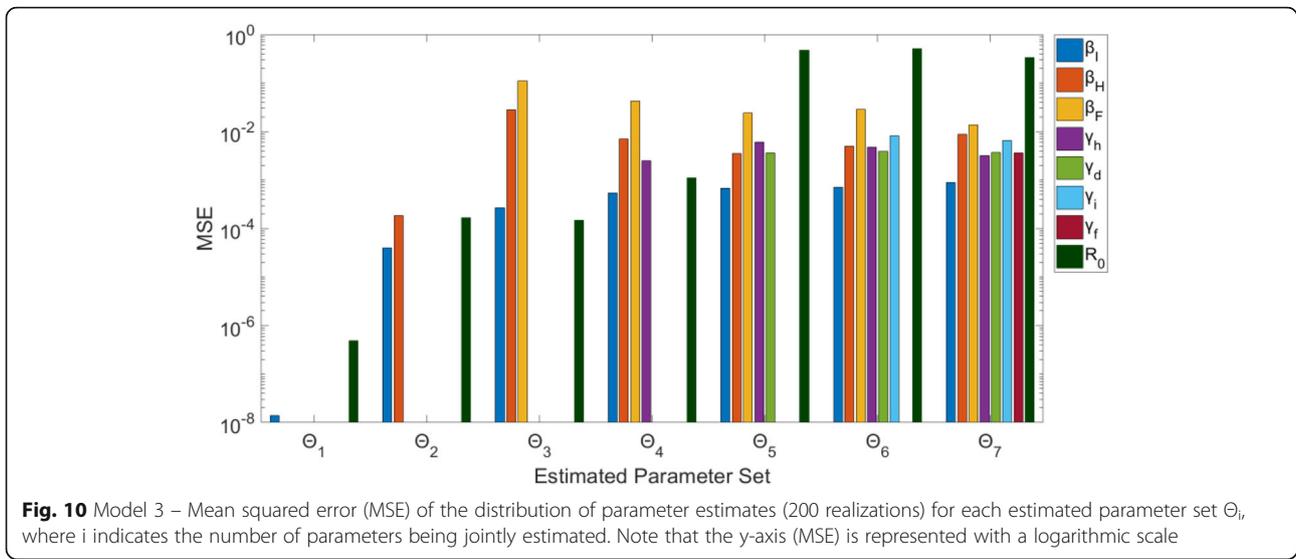


Fig. 9 Model 3–95% confidence intervals (vertical red lines) for the parameter estimate distributions obtained from the 200 realizations of the simulated datasets. Mean estimated parameter value is denoted by red x, and the true parameter value is represented by the blue horizontal line. Θ_i denotes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



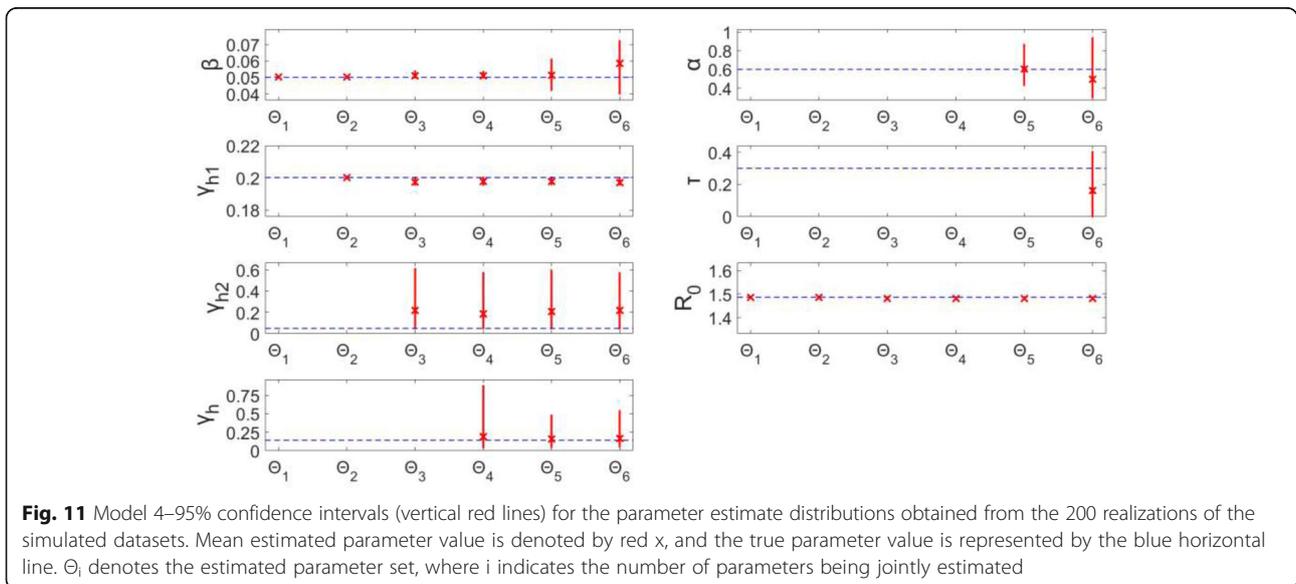
uncertainty: $(\Theta_6 - R_0: (1.480, 1.486))$). The R_0 estimates have little error, as $MSE < 10^{-4}$ for all Θ_i . This is consistent with the previous models in that R_0 estimates are robust to the uncertainty and bias of the other estimated parameters.

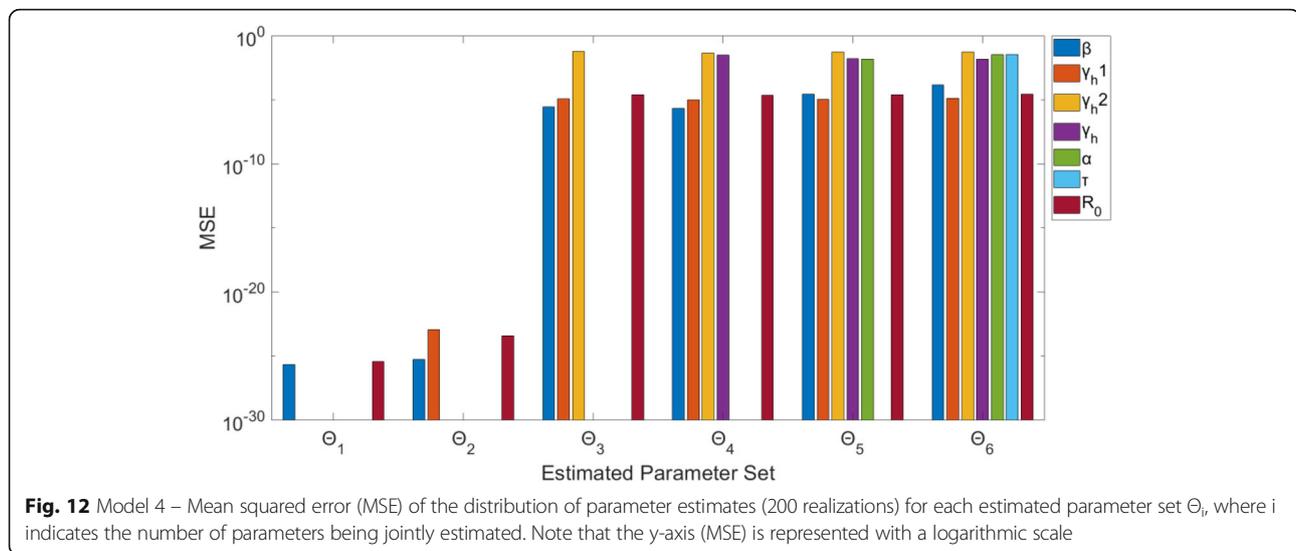
Discussion

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 (Fig. 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 (Figs. 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.

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.

Additional files

Additional file 1: 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. (TIF 5423 kb)

Additional file 2: 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. (TIF 5423 kb)

Additional file 3: 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 best fit. (TIF 5423 kb)

Acknowledgements

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

Funding

GC acknowledges financial support from the NSF grant 1414374 as part of the joint NSF-NIH-USDA Ecology and Evolution of Infectious Diseases program; UK Biotechnology and Biological Sciences Research Council grant BB/M008894/1.

Availability of data and materials

The datasets generated and/or analyzed in this study can be reproduced using the methods and Tables 1–4 or are available from the corresponding author on reasonable request. Matlab code is also available upon request.

Authors' contributions

KR and GC designed the study. KR analyzed the data and KR and GC interpreted the data. GC and KR contributed to further draft and edit the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

All of the data employed in this study were generated through simulations. Data are deemed exempt from institutional review board assessment.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

¹Department of Population Health Sciences, School of Public Health, Georgia State University, Atlanta, GA, USA. ²Division of International Epidemiology and Population Studies, Fogarty International Center, National Institute of Health, Bethesda, MD, USA.

Received: 16 August 2018 Accepted: 10 December 2018

Published online: 14 January 2019

References

- Anderson RM, May RM. Infectious diseases of humans: dynamics and control. Oxford: Oxford University Press; 1991.
- 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. *J Math Biol.* 1990;28(4):365–82.
- 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.
- 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. *J R Soc Interface.* 2010;7(43):271–83.
- 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.
- 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 Med.* 2017;15(1):42.

7. Banks HT, Holm K, Robbins D. Standard error computations for uncertainty quantification in inverse problems: asymptotic theory vs. bootstrapping. *Math Comput Model.* 2010;52:1610–25.
8. Gibson GJ, Stretzaris G, Thong D. Comparison and assessment of epidemic models. *Stat Sci.* 2018;33(1):19–33.
9. Banks H, Davidian M, Samuels J Jr, 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.* Dordrecht, 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. *Stat Sci.* 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. *Mech Syst Signal Process.* 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;25:89–100.
15. Miao H, Xia X, Perelson AS, Wu H. On identifiability of nonlinear ODE models and applications in viral dynamics. *SIAM Rev.* 2011;1:3.
16. Pohjanpalo H. System identifiability based on power-series expansion of solution. *Math Biosci.* 1978;41:21–33.
17. Eisenberg MC, Robertson SL, Tien JH. Identifiability and estimation of multiple transmission pathways in cholera and waterborne disease. *J Theor Biol.* 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. *Math Biosci.* 2002;180:29–48.
23. Chowell G, Nishiura H. Comparative estimation of the reproduction number for pandemic influenza from daily case notification data. *J R Soc 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. *Epidemiol Infect.* 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 Trans Biomed Eng.* 1976;BME-23(2):93.
30. Jacquez JA. *Compartmental analysis in biology and medicine.* 2nd ed. Ann Arbor: University of Michigan Press; 1985.
31. Raue A, Kreutz C, Maiwald T, Bachmann J, Schilling M, Klingmüller 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. *Front Microbiol.* 2015;6:257.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

