Mathematical Biosciences 299 (2018) 1–18

Contents lists available at ScienceDirect

Mathematical Biosciences

journal homepage: www.elsevier.com/locate/mbs

Structural and practical identifiability analysis of outbreak models

Necibe Tuncer[a], Trang T. Le*,[b]

[a] Department of Mathematical Sciences, Florida Atlantic University, Science Building, Room 234, 777 Glades Road, Boca Raton, FL 33431, USA
[b] Department of Mathematics, University of Tulsa, 800 S Tucker Drive, Tulsa, OK 74104, USA

ARTICLE INFO

Keywords:
Structural identifiability
Practical identifiability
Outbreak models
Emerging infectious disease
Cumulative incidence

AMS:
92D30
92D40

ABSTRACT

Estimating the reproduction number of an emerging infectious disease from an epidemiological data is becoming more essential in evaluating the current status of an outbreak. However, these studies are lacking the fundamental prerequisite in parameter estimation problem, namely the structural identifiability of the epidemic model, which determines the possibility of uniquely determining the model parameters from the epidemic data. In this paper, we perform both structural and practical identifiability analysis to classical epidemic models such as SIR (Susceptible-Infected-Recovered), SEIR (Susceptible-Exposed-Infected-Recovered) and an epidemic model with the treatment class (SITR). We performed structural identifiability analysis on these epidemic models using a differential algebra approach to investigate the well-posedness of the parameter estimation problem. Parameters of these models are estimated from different data types, namely prevalence, cumulative incidences and treated individuals. Furthermore, we carried out practical identifiability analysis on these models using Monte Carlo simulations and Fisher's Information Matrix. Our study shows that the SIR model is both structurally and practically identifiable from the prevalence data. It is also structurally identifiable to cumulative incidence observations, but due to high correlations of the parameters, it is practically unidentifiable from the cumulative incidence data. Furthermore, we found that none of these simple epidemic models are practically identifiable from the cumulative incidence data which is the standard type of epidemiological data provided by CDC or WHO. Our analysis with simple SIR model suggest that the health agencies, if possible, should report prevalence rather than incidence data.

1. Introduction

In recent years, using outbreak data to interpret the future of an emerging infection by means of mathematical models has gained significant attention. Examples of such model-based forecasts of an emerging pathogen are SARS [15,20], pandemic H1N1 Influenza [21], Cholera in Haiti [22] and most recently Ebola Virus in West Africa [7,11]. In early stages of an outbreak, it is crucial to specify some of the key factors of the outbreak such as transmission rate of the pathogen, the total outbreak size, the magnitude and the timing of the epidemic peak, duration of the incubation and infectiousness periods. These key factors determine an epidemiologically important threshold value called basic reproduction number, $\mathscr{R}_0$. The basic reproduction number, which is the average number of infections caused by one infected individual while being infectious in a totally susceptible population, determines whether the disease will die out or persist in the population.

An algebraic expression for $\mathscr{R}_0$ can be derived from the Ordinary Differential Equations (ODE) system modeling the emerging disease [16,25]. However many of the parameters of the basic reproduction number $\mathscr{R}_0$ cannot be determined using clinical data, since such data will be rare during the early stages of an outbreak. Hence, one uses indirect methods to estimate the parameters of mathematical model from the incidence reports provided by health agencies as the outbreak progresses. Such estimates are obtained by fitting mathematical models to the data. The only way to determine whether an emerging disease will spread among the population or will be controlled highly relies on interpreting the data to quantify the parameters of the model. The type of data available through government agencies such as World Health Organization (WHO) or Center for Disease Control (CDC) are not standardized across organizations and situations. Depending on the context, different reporting infrastructures often result in different types of data with a wide range of qualities. The prevalence, new incidences, cumulative number of incidences or deaths are some of the data types that are available for further analysis [6,26]. The type of data being used has also significant effect on the parameter estimation problem. Thus, it is crucial to understand the process of estimating the parameters of the epidemic model from given data.

A fundamental prerequisite for the parameter estimation problem to

* Corresponding author.
E-mail addresses: ntuncer@fau.edu (N. Tuncer), trang-le@utulsa.edu (T.T. Le).

https://doi.org/10.1016/j.mbs.2018.02.004
Received 30 May 2017; Received in revised form 6 February 2018; Accepted 17 February 2018
Available online 29 March 2018

be well posed is the *structural identifiability* of the mathematical model of the emerging disease. Structural identifiability studies whether the parameters of the model can be recovered from the observed state (output) under ideal conditions such as noise-free data and error-free model. The structural identifiability analysis can be done without any actual experimental data, hence it is also called the *prior* identifiability. It addresses the well-posedness of the parameter estimation problem under ideal situations, so it is a necessary but not a sufficient property to ensure the accurate identification of model parameters from the real noisy data. A model which is structurally identifiable may not be practically identifiable. On the other hand, if a model is structurally unidentifiable, then any parameter estimation obtained by a numerical optimization algorithm will be unreliable. A mathematical model which is structurally identifiable, might be unidentifiable in practice. Structural identifiability analysis relies on the assumptions that the model structures are accurate and there are no measurement errors, which are not valid in practice. Moreover, in real-world data, additional unknown parameters such as the reporting rate (fraction of cases reported) to be estimated and the lack of full time-course data from currently-evolving epidemics pose even more challenges in model parameter estimation. Therefore, even though the structural identifiability analysis concludes that the parameters of the model are uniquely identifiable, when noisy data are considered the parameter estimation problem might reveal unreliable results.

In this work, we study both structural and practical identifiability of several outbreak models (SIR, SEIR, and SITR) for different data types (prevalence, cumulative incidences and treated individuals). Structural identifiability of these simple infectious disease models have been studied extensively in the literature [5,9,10,17,23]. Structural identifiability of SIR model with seasonal forcing have been studied in [10] using the Lie derivatives approach, SIR model with demography and a cholera model have been studied in [9] using differential algebra approach. The purpose of this study is to study both structural and practical identifiability of simple outbreak models with different data types. For example, as stated in [9] the SIR model is structurally identifiable from both prevalence and incidence data. To ensure the reliability of the parameter estimation, we continue with the practical identifiability analysis of the epidemic models. We perform practical identifiability analyses to further analyze the well-posedness of the parameter estimation problem of the outbreak models. We found that all these simple models considered in this study, SIR, SEIR, and SITR are not practically identifiable from cumulative incidence data, which is the data type reported by health organizations. This study shows that cumulative incidences alone is not enough to identify the parameters of simple infectious disease models (see Table 23).

The paper is organized as follows: the outbreak models (SIR, SEIR, and SITR models) used in this study are introduced in Section 2. The structural identifiability analysis of these models are performed in Section 3. Section 4 summarizes the practical identifiability methods such as Monte Carlo simulations and Fishers Information Matrix used in this study. In Section 5, we perform numerical experiments with synthetic data to all three epidemiological models. The MATLAB code for this research has been made available at https://github.com/NecibeTuncer/Outbreak_Models. The use of synthetic data with known noise structure and level allows us to investigate the practical identifiability of the epidemic models for different data types. Furthermore, in addition to synthetic data, we also performed the practical identifiability of SEIR model with Ebola outbreak data in West Africa in 2014.

2. Epidemiological models

The goal of this study is to determine whether the parameters of the epidemiological model can be estimated from the given data. For this purpose we choose a selection of epidemiological models and data structures. Since we are interested in an emerging infection, we consider epidemiological models without any demography.

The first model is the simplest model structure of Susceptible-Infectious-Recovered model. The population, $N(t)$ is divided into three nonintersecting classes susceptible $S(t)$, infectious $I(t)$, and recovered $R(t)$. The transmission is described by the standard incidence, $\beta\frac{SI}{N}$, where β is the transmission rate. The infected individuals recover at rate α and move to the recovered class with full immunity. The first model takes the following form.

SIR Model:
$$\begin{cases} \dfrac{dS}{dt} &= -\beta\dfrac{SI}{N}, \\ \dfrac{dI}{dt} &= \beta\dfrac{SI}{N} - \alpha I, \\ \dfrac{dR}{dt} &= \alpha I, \end{cases} \tag{2.1}$$

which is equipped with initial conditions $S(0) = S_0,\ I(0) = I_0,\ R(0) = 0$. The total population $N(t) = S(t) + I(t) + R(t)$ satisfies the differential equation $N'(t) = 0$, hence $N(t) = N = S_0 + I_0$ is constant. The basic reproduction number for this SIR model is $\mathscr{R}_0 = \frac{\beta}{\alpha}$. We are interested in estimating the parameters $\mathbf{p} = [\beta, \alpha]$ for this model.

As a second model, we consider an epidemic model which involves compartments related to disease progression stages. In the above model (2.1), it is assumed the susceptible individuals moved to the infectious class immediately after infection. But for diseases such as influenza, infected individuals do not become infectious immediately, since the pathogen needs to replicate and reach a threshold value for host to become infectious. So, for the next model we add a latent, or exposed, class $E(t)$ to the SIR model (2.1). Let $\frac{1}{\eta}$ denote length of the latent period, the time for which an individual is infected but not yet infectious. The SEIR model takes the following form.

SEIR Model:
$$\begin{cases} \dfrac{dS}{dt} &= -\beta\dfrac{SI}{N}, \\ \dfrac{dE}{dt} &= \beta\dfrac{SI}{N} - \eta E, \\ \dfrac{dI}{dt} &= \eta E - \alpha I, \\ \dfrac{dR}{dt} &= \alpha I, \end{cases} \tag{2.2}$$

which is equipped with initial conditions $S(0) = S_0,\ E(0) = E_0,\ I(0) = 0,\ R(0) = 0$. The total population $N(t)$ is constant and the basic reproduction number is $\mathscr{R}_0 = \frac{\beta}{\alpha}$. For this model, we estimate the parameters $\mathbf{p} = [\beta, \eta, \alpha]$. SEIR (2.2) is one of the models that was used to predict the 2014 Ebola outbreak in West Africa [1]. In that study, basic reproduction number $\mathscr{R}_0$ is estimated by fitting the SEIR model (2.2) to the cumulative number of cases and deaths provided by the WHO [1].

For the next model we consider an epidemic model which incorporates the strategies applied for disease control. Some examples of measures taken to prevent and control infectious diseases involve quarantine, isolation, vaccination and treatment. We are going to take treatment for the case of epidemic models which includes control measures. Let $T(t)$ be the number of individuals in the treatment class. Suppose that fraction γ per unit time of infected individuals are selected for treatment and move to the treatment class. Individuals in the treatment class can infect susceptible individuals at a reduced transmission rate $\delta\beta$ where $0 \le \delta \le 1$. The treatment model takes the following form.

Treatment Model:

$$\begin{cases} \dfrac{dS}{dt} &= -\beta\dfrac{SI}{N} - \delta\beta\dfrac{ST}{N}, \\ \dfrac{dI}{dt} &= \beta\dfrac{SI}{N} + \delta\beta\dfrac{ST}{N} - (\alpha+\gamma)I, \\ \dfrac{dT}{dt} &= \gamma I - \nu T, \\ \dfrac{dR}{dt} &= \alpha I + \nu T, \end{cases} \tag{2.3}$$

which is equipped with the initial conditions $S(0) = S_0,\ I(0) = I_0,\ T(0) = 0,\ R(0) = 0$. The total population size is constant and $(S + T + I)' = -\alpha I - \nu T$. Thus, $\lim_{t\to\infty} S(t) = S_\infty > 0$ and $\lim_{t\to\infty} I(t) = 0$ and $\lim_{t\to\infty} T(t) = 0$. For this model we estimate the parameters $\mathbf{p} = [\beta, \delta, \alpha, \gamma, \nu]$. The basic reproduction number for the treatment model (2.3) is

$$\mathscr{R}_0 = \frac{\beta\nu + \gamma\beta\delta}{\nu(\alpha+\gamma)}.$$

3. Structural identifiability

Parameters of the basic reproduction number $\mathscr{R}_0$ which characterizes the future of an epidemic usually cannot be measured directly. Quantification of epidemiological parameters such as transmission rate, disease induced death rate and recovery rate are usually approached indirectly via parameter estimation problem in the context of least squares.

Without loss of generality, we will represent the epidemiological models (2.1)–(2.3) in the following compact form,

$$\frac{d\mathbf{x}}{dt} = f(\mathbf{x}(t), \mathbf{p}) \quad \mathbf{x}(0) = \mathbf{x}_0 \tag{3.1}$$

where $\mathbf{x}(t)$ denotes the state variables of the epidemic model, and $\mathbf{p}$ denotes the vector of epidemiological parameters such as β, α or η depending on the model. In this study, we assume that initial state variables, $\mathbf{x}_0$ are known and suppose that observations are given by the output function $g(\mathbf{x}(t), \mathbf{p})$. The observation in epidemiology can be prevalence, incidences, cumulative number of cases, or cumulative number of deaths. The observations $\{y_i\}_{i=1}^n$ are obtained at discrete time points $t_1,\ t_2,\ \ldots t_n$. We then define the statistical model by [2],

$$y_i = g(\mathbf{x}(t_i), \hat{\mathbf{p}}) + E_i \tag{3.2}$$

where $\hat{\mathbf{p}}$ denotes the true parameters that generate the observations $\{y_i\}_{i=1}^n$ and E_i are the random variables that represent the observation or measurement error which cause the observations not fall exactly on the points $g(\mathbf{x}(t_i), \hat{\mathbf{p}})$ of the smooth path $g(\mathbf{x}(t), \hat{\mathbf{p}})$. In a general setting, the measurements errors are assumed to have the following form,

$$E_i = g(\mathbf{x}(t_i), \hat{\mathbf{p}})^{\xi}\epsilon_i \tag{3.3}$$

where $\xi \geq 0$ and ϵ_i are independent and identically distributed with mean zero and constant variance σ_0^2. The random variables y_i have mean $\mathbb{E}(y_i) = g(\mathbf{x}(t_i), \hat{\mathbf{p}})$ and variances $Var(y_i) = g(\mathbf{x}(t_i), \hat{\mathbf{p}})^{2\xi}\sigma_0^2$. Varying ξ allows for varying error scales in the measurements.

For the parameter estimation problem, we suppose that the emerging infectious disease is exactly described by one of the deterministic models (2.1)–(2.3), that is there is no modeling error and the expected value of the random variables ϵ_i are zero, hence $\mathbb{E}(\epsilon_i) = 0$.

The parameter estimation in terms of likelihood approach can be stated as, finding the parameter set $\mathbf{p}$ that maximizes the likelihood of observing the given data. Maximum likelihood is a general approach for estimating model parameters. Let likelihood, that is the probability of observing the data given the model parameters be denoted by $L(y|\mathbf{p})$, then

$$\hat{\mathbf{p}} = \underset{\mathbf{p}}{\operatorname{argmax}}\, L(\mathbf{y}|\mathbf{p}). \tag{3.4}$$

If in the statistical model (3.2), (3.3), $\xi = 0$, then the observed data y_i will have Gaussian distribution with mean $g(\mathbf{x}(t_i), \hat{\mathbf{p}})$ and constant variance σ_0^2. In this case, the maximum likelihood estimate is equivalent to least square estimate, that is

$$\hat{\mathbf{p}} = \min_{\mathbf{p}} \sum_{i=1}^{n} (y_i - g(\mathbf{x}(t_i), \mathbf{p}))^2. \tag{3.5}$$

We use the *relative error* model, that is $\xi = 1$ in (3.3) and use *ordinary least squares*in the parameter estimation problem. We also performed weighted least squares but did not obtain different results; hence, the results are not shown.

Structural identifiability concerns with the possibility of uniquely determining the epidemiological parameters from the given output function $g(\mathbf{x}(t), \mathbf{p})$, under ideal conditions such as noise-free data and error-free model. If the epidemiological model reveals all its parameters uniquely from the given output function, then the model (or its parameters) is said to be structurally globally identifiable. If all the parameters of the epidemiological model have finitely many solutions for the given output, then the model (or its parameters) is said to be structurally locally identifiable. If some parameters of the model yield infinitely many solutions for the parameter estimation problem, then the model is said to be unidentifiable. Even if an epidemiological model is structurally identifiable, this does not guarantee an accurate estimation of model parameters from real data. But, if the epidemiological model is not structurally identifiable then any parameter estimation obtained by some numerical optimization algorithm will be totally unreliable and random. It is crucial to check the identifiability of the epidemic model before performing any parameter estimation experiment which may not give any accurate information about the future of an epidemic. The definition of the structural identifiability is given as [18].

Definition 1. A parameter set $\mathbf{p}$ is called *structurally globally (or uniquely) identifiable* if for every $\mathbf{q}$ in the parameter space, the equation

$$g(\mathbf{x}(t), \mathbf{p}) = g(\mathbf{x}(t), \mathbf{q}) \Rightarrow \mathbf{p} = \mathbf{q}.$$

That is any unequal parameter set yield different observations and hence the corresponding noise-free data are as well distinct.

Definition 2. Let $\mathscr{N}(\mathbf{p})$ denote the neighborhood of the parameter $\mathbf{p}$, then $\mathbf{p}$ is called *structurally locally identifiable* if for every $\mathbf{q}$ in the parameter space, the equation

$$g(\mathbf{x}(t), \mathbf{p}) = g(\mathbf{x}(t), \mathbf{q}) \quad \text{and} \quad \mathbf{p} \in \mathscr{N}(\mathbf{p}) \Rightarrow \mathbf{p} = \mathbf{q}.$$

There are many mathematical methods to study the structural identifiability of dynamic system ODE models. Some of these methods are differential algebra approach, Taylor series approach, generating series approach and a method based on the Implicit Function Theorem [3,9,14,18,19]. For details about these methods, we refer the reader to the good reviews written on the topic [8,18]. In this paper, we will use differential algebra approach to study the structural identifiability of the epidemiological models (2.1)–(2.3). We study whether an epidemiological model is structured to reveal its parameters from observations of prevalence and cumulative number of cases using the differential algebra approach. We will briefly summarize the differential algebra approach here, but for more details reader is referred to [3,18].

The first step in differential algebra approach is to find the input-output equations of the dynamical ODE model. This is done by reducing the model to its *characteristic set* via Ritt's algorithm. The characteristic set is then used to derive the input-output equations which contains all the identifiability information of the model. Input-output equations simply put are a subset of the characteristic set. The characteristic set of a dynamical ODE model is not unique, hence it depends on the chosen ranking of the state variables. In order to obtain the desired input-output equation, the observed state variable should be ranked the smallest. However, the coefficients of the input-output equation can be fixed uniquely by normalizing the equations to make them monic.

For instance, consider the SIR model (2.1) and take prevalence to be the data set provided for the parameter estimation problem. Although it is well known that the SIR model (2.1) is structurally identifiable to prevalence data, we consider this model here as a simple example to demonstrate how the differential algebra approach can be carried out to assess a model's identifiability. Since $N = S_0 + I_0$, we are only interested in estimating the parameters $\mathbf{p} = [\beta, \alpha]$. Hence, $g(\mathbf{x}(t), \mathbf{p}) = I(t, \mathbf{p})$, then an appropriate choice for ranking the state variables would be as $I < S < R$. The ranking $I < S < R$ yields to the following characteristic equations

$$\begin{aligned} NI^{''}I - NI^2 + \beta I^{'}I^2 + \alpha\beta I^3 &= 0,\\ \beta SI - NI^{'} - \alpha NI &= 0,\\ R^{'} - \alpha I &= 0. \end{aligned} \tag{3.6}$$

Then the normalized input-output equation for the prevalence data is

$$I^{''}I - I^2 + \frac{\beta}{N}I^{'}I^2 + \frac{\alpha\beta}{N}I^3 = 0. \tag{3.7}$$

The input-output equation is a differential algebraic equation of the output, which in this case is the prevalence, $I(t)$ and its derivatives, with coefficients being the parameters of the model. The infected number of individuals obtained by solving the model (2.1) is the same as the solution of the input-output equation (3.7). Within the differential algebra approach, the ODE model is said to be structurally identifiable if the map from the parameter space to the coefficients of the input-output equations is injective [9]. Thus, the definition of the structural identifiability becomes:

Definition 3. The dynamical model is structurally globally identifiable if and only if

$$c(\mathbf{p}) = c(\mathbf{q}) \Rightarrow \mathbf{p} = \mathbf{q}$$

where $c(\mathbf{p})$ are the coefficients of the normalized input-output equation.

Suppose to the contrary that there exist another set of parameters $\mathbf{q} = [\beta_q, \alpha_q]$ which produced the same output, then the coefficients of the input-output equation yield to the following

$$\beta = \beta_q, \qquad \text{and} \qquad \alpha\beta = \alpha_q\beta_q,$$

which clearly states that $\beta = \beta_q$, $\alpha = \alpha_q$. Hence the SIR model (2.1) is structurally identifiable to prevalence data.

As a second example, suppose that the observations are given as the cumulative incidences. Since S_0 is given, we have

$$g(\mathbf{x}(t), \mathbf{p}) = \int_0^t \beta\frac{S(\tau)I(\tau)}{N}d\tau = S_0 - S(t).$$

Thus, choosing the ranking as $S < I < R$ would yield to the following characteristic equation

$$\begin{aligned} NSS^{''} - N(S^{'})^2 - \beta S^{'}S^2 + \alpha NS^{'}S &= 0,\\ \beta SI + NS^{'} &= 0,\\ \beta SR^{'} + N\alpha S^{'} &= 0. \end{aligned} \tag{3.8}$$

The normalized input-output equation for the cumulative incidences is

$$SS^{''} - (S^{'})^2 - \frac{\beta}{N}S^{'}S^2 + \alpha S^{'}S = 0.$$

Clearly by Definition 2, we conclude that SIR model (2.1) is structurally identifiable to cumulative number of cases data. Thus, we have established the following result.

Proposition 1. *The SIR Model (2.1) is structurally identifiable to the prevalence and cumulative incidence observations.*

Next, we prove that the epidemiological model with the latent class (2.2) is not globally identifiable from both prevalence and cumulative incidences observations, but only locally identifiable.

Proposition 2. *The SEIR Model (2.2) is locally structurally identifiable to prevalence observations.*

Proof. The normalized input-output equation derived by DAISY [3] is

$$I^{''^{'}}I - I^{''}I^{'} + \frac{\beta}{N}I^{''}I^2 + (\alpha+\eta)I^{''}I - (\alpha+\eta)(I^{'})^2 + \frac{\alpha\beta+\beta\eta}{N}I^{'}I^2 + \frac{\alpha\beta\eta}{N}I^3 = 0 \tag{3.9}$$

Note that N can be determined by the initial conditions. Using Definition 3, we set $c(\mathbf{p}) = c(\mathbf{q})$, where $\mathbf{p} = [\beta, \alpha, \eta]$ and $\mathbf{q} = [\beta_q, \alpha_q, \eta_q]$ and obtain

$$\beta = \beta_q, \qquad \alpha + \eta = \alpha_q + \eta_q, \qquad \alpha\beta + \beta\eta = \alpha_q\beta_q + \beta_q\eta_q, \qquad \alpha\beta\eta = \alpha_q\beta_q\eta_q \tag{3.10}$$

we then solve these equations (3.10) using `GroebnerBasis` in Mathematica and get two sets of solutions.

$$\{\beta = \beta_q, \alpha = \alpha_q, \eta = \eta_q\}, \qquad \{\beta = \beta_q, \alpha = \eta_q, \eta = \alpha_q\} \tag{3.11}$$

Hence, the SEIR model (2.2) is only locally identifiable. □

Remark 1. Note that, if α is fixed (or alternatively η is fixed), then the SEIR model (2.2), would be globally (uniquely) identifiable from prevalence observations.

The cumulative incidence observations for the SEIR model (2.2) is given by

$$g(\mathbf{x}(t), \mathbf{p}) = \int_0^t \eta E(\tau)d\tau.$$

Proposition 3. *The SEIR Model (2.2) is structurally identifiable to cumulative incidence observations.*

Proof. Let the cumulative incidences $C(t) = \int_0^t \eta E(\tau)d\tau$ be an additional state variable in the SEIR model (2.2). Hence, augmenting (2.2) with $C'(t) = \eta E(t)$, we obtain the normalized input-output relations using DAISY and then set $c(\mathbf{p}) = c(\mathbf{q})$. We solve the system $c(\mathbf{p}) = c(\mathbf{q})$ using `GroebnerBasis` in Mathematica and obtain,

$$\{\beta = \beta_q, \alpha = \alpha_q, \eta = \eta_q\}. \tag{3.12}$$

Hence, the SEIR model (2.2) is structurally identifiable to cumulative incidences observations. □

Proposition 4. *The parameters of the treatment model (2.3) cannot be identified from the observations of cumulative incidences.*

Proof. The normalized input-output equation for cumulative incidences is

$$\begin{aligned} &-3S^{''}S^{'}S - \frac{\beta}{N}S^{''}S^3 + (\alpha+\gamma+\nu)S^{''}S^2 + 2(S^{'})^3 - (\alpha+\gamma+\nu)(S^{'})^2S\\ &\quad - 3\frac{\beta\delta\gamma+\beta\nu}{N}S^{'}S^3 + (\alpha\nu+\gamma\nu)S^{'}S^2 = 0 \end{aligned} \tag{3.13}$$

Setting $c(\mathbf{p}) = c(\mathbf{q})$, we get the following two sets of infinitely many solutions.

$$\begin{aligned} &\{\beta = \beta_q,\ \alpha + \gamma = \alpha_q + \gamma_q,\ \nu = \nu_q,\ \delta\gamma = \delta_q\gamma_q\},\\ &\{\beta = \beta_q, \alpha + \gamma = \nu_q,\ \nu = \alpha_q + \gamma_q,\ \nu + \delta\gamma = \nu_q + \delta_q\gamma_q\} \end{aligned} \tag{3.14}$$

So, there exists infinitely many solutions. Hence the treatment model is not structured to identify its parameters from cumulative incidences observations. □

Remark 2.

(i) If the parameter α is fixed to its true value, that is suppose that

$\alpha = \alpha_q$ is given, then solving the system $c(\mathbf{p}) = c(\mathbf{q})$, results in the following,

$$\{\beta = \beta_q,\ \alpha = \alpha_q,\ \gamma = \gamma_q,\ \nu = \nu_q,\ \delta = \delta_q\},$$
$$\{\beta = \beta_q,\ \alpha = \alpha_q,\ \gamma = \gamma_q,\ \nu = \alpha_q + \gamma_q,\ \nu + \delta\gamma = \nu_q + \delta_q\gamma_q\} \tag{3.15}$$

That is we get two sets of solutions. Fixing $\alpha = \alpha_q$ yields locally structurally identifiable model.

(ii) Furthermore, if both parameters, α and δ are fixed to true values, then the treatment model becomes globally (uniquely) identifiable.

Proposition 5. *The parameters of the treatment model (2.3) cannot be identified from observations of the treatment state variable.*

Proof. We first derive the normalized input-output equation in DAISY [3], then setting, $c(\mathbf{p}) = c(\mathbf{q})$, we get the following nonlinear system.

$$
\begin{aligned}
&(\alpha\beta + \beta\gamma + \beta\nu)\gamma_q = (\alpha_q\beta_q + \beta_q\gamma_q + \beta_q\nu_q)\gamma,\\
&\beta\gamma_q = \beta_q\gamma,\\
&(2\alpha\beta\delta\gamma + 2\beta\delta\gamma^2 + 2\beta\delta\gamma\nu + 3\alpha\beta\nu + 3\beta\gamma\nu + 2\beta\nu^2)\gamma_q\\
&= \left(2\alpha_q\beta_q\delta_q\gamma_q + 2\beta_q\delta_q\gamma_q^2 + 2\beta_q\delta_q\gamma_q\nu_q + 3\alpha_q\beta_q\nu_q + 3\beta_q\gamma_q\nu_q + 2\beta_q\nu_q^2\right)\gamma,\\
&(\alpha\delta\gamma^2 + \delta\gamma^3 + \delta\gamma^2\nu + \gamma\nu^2)\gamma_q = \left(\alpha_q\delta_q\gamma_q^2 + \delta_q\gamma_q^3 + \delta_q\gamma_q^2\nu_q + \gamma_q\nu_q^2\right)\gamma,\\
&(2\beta\delta\gamma + 2\beta\nu)\gamma_q = (2\beta_q\delta_q\gamma_q + 2\beta_q\nu_q)\gamma,\\
&(\delta\gamma^2 + \gamma\nu)\gamma_q = \left(\delta_q\gamma_q^2 + \gamma_q\nu_q\right)\gamma,\\
&(\alpha\beta\delta^2\gamma^2 + \beta\delta^2\gamma^3 + \beta\delta^2\gamma^2\nu + 4\alpha\beta\delta\gamma\nu + 4\beta\delta\gamma^2\nu + 2\beta\delta\gamma\nu^2 + 3\alpha\beta\nu^2\\
&\quad + 3\beta\gamma\nu^2 + \beta\nu^3)\gamma_q\\
&= \left(\alpha_q\beta_q\delta_q^2\gamma_q^2 + \beta_q\delta_q^2\gamma_q^3 + \beta_q\delta_q^2\gamma_q^2\nu_q + 4\alpha_q\beta_q\delta_q\gamma_q\nu_q + 4\beta_q\delta_q\gamma_q^2\nu_q\right.\\
&\quad \left. + 2\beta_q\delta_q\gamma_q\nu_q^2 + 3\alpha_q\beta_q\nu_q^2 + 3\beta_q\gamma_q\nu_q^2 + \beta_q\nu_q^3\right)\gamma,\\
&(\beta\delta^2\gamma^2 + 2\beta\delta\gamma\nu + \beta\nu^2)\gamma_q = \left(\beta_q\delta_q^2\gamma_q^2 + 2\beta_q\delta_q\gamma_q\nu_q + \beta_q\nu_q^2\right)\gamma,\\
&(\alpha\delta\gamma^2 + \delta\gamma^3 + \delta\gamma^2\nu + \gamma\nu^2)\gamma_q = \left(\alpha_q\delta_q\gamma_q^2 + \delta_q\gamma_q^3 + \delta_q\gamma_q^2\nu_q + \gamma_q\nu_q^2\right)\gamma,\\
&(\alpha\beta\delta^2\gamma^2\nu + \beta\delta^2\gamma^3\nu + 2\alpha\beta\delta\gamma\nu^2 + 2\beta\delta\gamma^2\nu^2 + \alpha\beta\nu^3 + \beta\gamma\nu^3)\gamma_q\\
&= \left(\alpha_q\beta_q\delta_q^2\gamma_q^2\nu_q + \beta_q\delta_q^2\gamma_q^3\nu_q + 2\alpha_q\beta_q\delta_q\gamma_q\nu_q^2 + 2\beta_q\delta_q\gamma_q^2\nu_q^2 + \alpha_q\beta_q\nu_q^3\right.\\
&\quad \left. + \beta_q\gamma_q\nu_q^3\right)\gamma
\end{aligned}
\tag{3.16}
$$

We solve the system (3.16) using `GroebnerBasis` in Mathematica and obtain the following sets of infinitely many solutions.

$$
\begin{aligned}
&\{\beta = 0,\ \gamma = 0\}\\
&\left\{\frac{\beta}{\gamma} = \frac{\beta_q}{\gamma_q},\ \delta\gamma + \nu = \delta_q\gamma_q + \nu_q,\ \gamma + \alpha = \nu_q,\ \nu = \gamma_q + \alpha_q\right\},\\
&\{\beta\delta = \beta_q\delta_q,\ \gamma\delta = \gamma_q\delta_q,\ \alpha + \gamma = \alpha_q + \gamma_q,\ \nu = \nu_q\}
\end{aligned}
\tag{3.17}
$$

Thus, the treatment model is not structured to identify its parameters from observations of the treatment variable. □

Remark 3.

(i) Adding $\delta = \delta_q$ and $\gamma = \gamma_q$ to the system (3.16) yields unique solution. Thus, if the parameters, δ and γ are fixed, then the treatment model (2.3) becomes structurally identifiable.

(ii) The parameters of the treatment model (2.3) can not be identified from the observations of cumulative incidences or treatments. One way to obtain a well posed parameter estimation problem is to use multiple data sets. If both treatment and cumulative incidence data sets are considered then the treatment model (2.3) becomes locally identifiable (results not shown). Furthermore, in addition to using both data sets, if the parameter ν is fixed, then the parameters of the treatment model can be uniquely identified.

4. Practical identifiability

The structural identifiability is a characteristics of the model structure for the given output and it relies on the assumption that the model is error free and the output is noise free. But, practical identifiability depends not only on the model structure but also on the quantity and quality of the data together with the numerical optimization algorithm used for the parameter estimation problem (3.5). If an epidemiological model is structurally non-identifiable, then clearly it is practically non-identifiable as well. On the other hand, a model which is structurally identifiable may be practically non-identifiable. So, it is crucial to investigate whether structurally identifiable parameters can be estimated with a desirable accuracy from noisy data by implementing an optimization algorithm. In this study, we perform both structural and practical identifiability analysis for the outbreak models (2.1)–(2.3). For practical identifiability, we carry out both *Monte Carlo Simulations* and *Sensitivity Analysis*. While the *Monte Carlo Simulations* method helps us rigorously define a model's practical identifiability (Definition 4), *Sensitivity Analysis* provides the correlation structure of the parameters that guides the later parameter fixing procedure when a model is unidentifiable.

4.1. Monte Carlo simulation

Monte Carlo simulations have been widely used for practical identifiability of ODE models [18]. We perform Monte Carlo simulations by generating 1000 synthetic data sets using the true parameter set $\hat{\mathbf{p}}$ and adding noise at increasing levels. The Monte Carlo Simulations we performed are outlined in the following steps.

(1.) Solve the epidemiological model numerically with the true parameters $\hat{\mathbf{p}}$ and obtain the output vector $\mathbf{g}(\mathbf{x}(t), \hat{\mathbf{p}})$ at the discrete data time points $\{t_i\}_{i=1}^{n}$.

(2.) Generate $M = 1000$ data sets from the statistical model (3.2) with a given measurement error. Data sets are drawn from a normal distribution whose mean is the output vector obtained in step (1.) and standard deviation is the σ_0% of the mean. That is, we set $\xi = 1$ in the error structure given in (3.2)

$$y_i = g(\mathbf{x}(t_i), \hat{\mathbf{p}}) + g(\mathbf{x}(t_i), \hat{\mathbf{p}})\epsilon_i \quad i = 1, 2, \ldots, n$$

where $\mathbb{E}(\epsilon_i) = 0$ and $\mathrm{Var}(\epsilon_i) = \sigma_0^2$. Hence, the random variables y_i have mean $\mathbb{E}(y_i) = g(\mathbf{x}(t_i), \hat{\mathbf{p}})$ and variances $\mathrm{Var}(y_i) = g(\mathbf{x}(t_i), \hat{\mathbf{p}})^2\sigma_0^2$.

(3.) Fit the dynamical model $\mathbf{x}' = f(\mathbf{x}, t, \mathbf{p}) \quad \mathbf{x}(0) = \mathbf{x}_0$ to each of the M simulated data sets to estimate the parameter set $\mathbf{p}_j$ for $j = 1, 2, \ldots, M$. That is

$$\mathbf{p}_j = \min_{\mathbf{p}} \sum_{i=1}^{n} (y_i - g(\mathbf{x}(t_i), \mathbf{p}))^2, \quad j = 1, 2, \ldots, M.$$

Store the parameter estimate $\mathbf{p}_j$ in the matrix $\mathbf{Q}$ of dimensions $M \times s$ where s is the size of the vector $\mathbf{p}_j$.

(4.) Calculate the average relative estimation error for each parameter in the set $\mathbf{p}$ by [18]

$$ARE(p^{(k)}) = 100\% \frac{1}{M} \sum_{j=1}^{M} \frac{\left|\hat{p}^{(k)} - p_j^{(k)}\right|}{\hat{p}^{(k)}}$$

where $p^{(k)}$ is the kth parameter in the set $\mathbf{p}$, $\hat{p}^{(k)}$ is the kth parameter in the true parameter set $\hat{\mathbf{p}}$, and $p_j^{(k)}$ is the kth parameter in the set $\mathbf{p}_j$.

(5.) Calculate the mean $\mathbb{E}(\mathbf{p})$, standard deviation $SE(\mathbf{p})$ and co-variances $COV(\mathbf{p})$ of the parameter estimates using the following formulas.

$$\begin{aligned} \mathbb{E}(\mathbf{p}) &= \frac{1}{M}\sum_{j=1}^{M}\mathbf{p}_j \\ COV(\mathbf{p}) &= \frac{1}{M-1}\sum_{j=1}^{M}(\mathbf{p}_j - \mathbb{E}(\mathbf{p}))^T(\mathbf{p}_j - \mathbb{E}(\mathbf{p})) \\ SE(\mathbf{p}) &= \sqrt{COV(\mathbf{p})_{kk}} \quad k = 1, 2, \ldots, s \end{aligned} \tag{4.1}$$

(6.) We compute the correlation coefficients as,

$$\rho_{k,s} = \frac{COV(p^k, p^s)}{\sqrt{\mathrm{Var}(p^k)\mathrm{Var}(p^s)}}. \tag{4.2}$$

It is critical to note that if the correlation between two parameters is close to one, these two parameters are not distinguishable.

(7.) Repeat steps 1 through 5 with increasing level of noise, that is take $\sigma_0 = 0, 1, 5, 10, 20, 30\%$.

The computed *ARE*s give an insight about the identifiability of the parameters of the epidemiological model. If the epidemiological model is structurally identifiable, then when $\sigma_0 = 0$, that is when there are no noise in the data, the *ARE*s should be 0 or very close to 0. As expected, increasing the noise level in the data will increase the *ARE*s. But, if a parameter is not practically identifiable, then the *ARE* of that parameter will be significantly high even for a reasonable level of measurement error. Some of the parameters will be very sensitive to the noise in the data, and increasing the measurement errors will result in significantly high *ARE*s, then we say that the parameter is practically unidentifiable. We specifically define *practical identifiability* as follows:

Definition 4. If the *ARE* of a parameter is smaller than the measurement error σ_0, we say that the parameter is practically identifiable. A model is practically identifiable when all of its parameters are practically identifiable.

Monte Carlo simulations will reveal which parameters are practically unidentifiable, to understand the cause of the undeintifiability of the parameters, we further do the sensitivity analysis as explained in the next section.

4.2. Correlation matrix

Note that the parameter estimates $\mathbf{p}_j\, j = 1, 2, \ldots, M$ are random vectors, since $\epsilon_i = y_i - g(\mathbf{x}(t_i), \mathbf{p})$ are random variables. We suppose that $\mathbf{p}_j$ is a multivariate random vector which is normally distributed with mean as the true parameter $\hat{\mathbf{p}}$ and variance-covariance matrix $\hat{\Sigma}$, that is

$$\mathbf{p}_j \sim \mathcal{N}(\hat{\mathbf{p}}, \hat{\Sigma}).$$

We use the asymptotic theory as explained in [2] and suppose that

$$\mathbf{p}_j \sim \mathcal{N}(\hat{\mathbf{p}}, \hat{\Sigma}) \approx \mathcal{N}(\hat{\mathbf{p}}, \hat{\sigma}(F^n(\hat{\mathbf{p}})^T F^n(\hat{\mathbf{p}}))^{-1})$$

where $\hat{\sigma}$ is the standard deviation of the noise, and $F^n(\hat{\mathbf{p}})$ is the sensitivity function defined as

$$F^n(\hat{\mathbf{p}})_{ik} = \frac{\partial g}{\partial p^{(k)}}(\mathbf{x}(t_i), \hat{\mathbf{p}}) \quad i = 1, 2, \ldots, n \quad k = 1, 2, \ldots, s.$$

To calculate these sensitivities of the state variables with respect to the model parameters and initial conditions, we solve the following sensitivity equations simultaneously with the original system:

$$\frac{d}{dt}\frac{\partial \mathbf{x}}{\partial p^{(k)}} = \frac{\partial f}{\partial \mathbf{x}}\frac{\partial \mathbf{x}}{\partial p^{(k)}} + \frac{\partial f}{\partial p^{(k)}}$$

with initial conditions

$$\frac{\partial \mathbf{x}}{\partial p^{(k)}} = \mathbf{0}.$$

Hence, the inverse of the *Fisher Information Matrix* $(F^n(\hat{\mathbf{p}})^T F^n(\hat{\mathbf{p}}))^{-1}$ provides an approximation for the covariance matrix $\hat{\Sigma}$. More precisely,

$$\hat{\Sigma} \approx \Sigma^n = \hat{\sigma}^2(F^n(\hat{\mathbf{p}})^T F^n(\hat{\mathbf{p}}))^{-1}. \tag{4.3}$$

We then compute the correlation matrix, χ, which is based on the Fisher Information Matrix [12,24]

$$\chi_{ij} = \frac{\hat{\Sigma}_{ij}}{\sqrt{\hat{\Sigma}_{ii}\hat{\Sigma}_{jj}}} \quad i \neq j, \qquad \chi_{ij} = 1, \quad i = j \tag{4.4}$$

The components of the correlation matrix gives the correlation coefficient between parameter estimates, namely χ_{ij} gives the correlation coefficient between the parameters $\hat{p}^{(i)}$ and $\hat{p}^{(j)}$. If χ_{ij} is close to 1, then the parameters $\hat{p}^{(i)}$ and $\hat{p}^{(j)}$ are strongly correlated, that is one parameter depends on the other parameter and these two parameters cannot be estimated uniquely. That is if the correlation coefficient χ_{ij} is close to 1, then parameters are said to be practically unidentifiable.

5. Numerical experiments

In order to fully assess the performance of the estimation process, we generate data from selected models with known parameters and noise structure. Multiple simulated data sets at different noise levels help us directly examine the uncertainty in the parameter estimates. Moreover, synthetic data allow us to exactly calculate quantities such as the variance-covariance matrix instead of having to use estimated values.

We consider a population size of 101, in which 1 population unit is initially infectious, and the rest are susceptible. We use `MATLAB ode45` routine to solve for the epidemiological models, starting from $t = 0$ and ending at t_n, yielding output at $n + 1$ evenly spaced time points. The duration of the outbreak t_n is taken to be the first time at which the infected population $I(t)$ falls below 1. By doing this, we hope to accurately capture the epidemic timescale. To estimate the parameters from given models, we use Nelder–Mead algorithm implemented in `MATLAB fminsearchbnd`.

5.1. SIR model

5.1.1. Experiment A: prevalence data

Using differential algebra analysis, we showed that the SIR model (2.1) is structurally identifiable to prevalence observations. For the practical identifiability analysis, we first perform Monte Carlo simulations as described in the preceding section. We take $[\hat{\beta}, \hat{\alpha}] = [4, 1]$ as the true parameter set, and generate $M = 1000$ replicates of prevalence data $y_i = I(t_i, \hat{\mathbf{p}}) + I(t_i, \hat{\mathbf{p}})\epsilon_i$, at $n = 51$ time points, where $\mathbb{E}(\epsilon_i) = 0$ and $\mathrm{Var}(\epsilon_i) = \sigma_0^2$. Using `fminsearchbnd`, we fit the SIR model to each of the 1000 replicate data sets with the starting guess of $\beta_0 = \alpha_0 = 0.1$. The computed average relative estimation errors (AREs) at six measurement error levels, $\sigma_0 = 0, 1, 5, 10, 20,$ and 30% of parameters β and α are displayed in Table 3.

When there is no measurement error ($\sigma_0 = 0\%$), both parameters are well identified (max ARE is of order 10^{-7}), which verifies the theoretical identifiability result. This also suggests that the numerical method, Nelder–Mead algorithm [13] suited well for the parameter estimation problem since the parameter estimates converge to the true values when the sample size is large enough and the measurement error is small enough. As the error level increases, both parameters' AREs increase reasonably. Even when the error level reaches 30%, both AREs

Table 1
Definition of the variables in the epidemiological models.

Variable	Meaning
$N(t)$	Total population size
$S(t)$	Number of susceptible individuals
$I(t)$	Number of infected individuals
$R(t)$	Number of recovered individuals
$E(t)$	Number of latent individuals
$T(t)$	Number of treated individuals

Table 2
Definition of the parameters in the epidemiological models.

Parameter	Meaning
β	Transmission rate
α	Recovery rate
η	Per capita rate of becoming infectious for latent individuals
γ	Fraction of infected individuals selected for treatment
δ	Reduced factor transmission rate of individuals in treatment class
ν	Recovery rate of individuals in treatment class

Table 3
Monte Carlo simulation results: Average relative estimation error (ARE) for parameters of the SIR model (2.1) when fitted to prevelance data.

Error level (%)	β(ARE%)	α(ARE%)
0	2.61E−07	1.13E−07
1	0.102	0.170
5	0.497	0.865
10	1.011	1.802
20	2.059	3.468
30	3.202	5.241

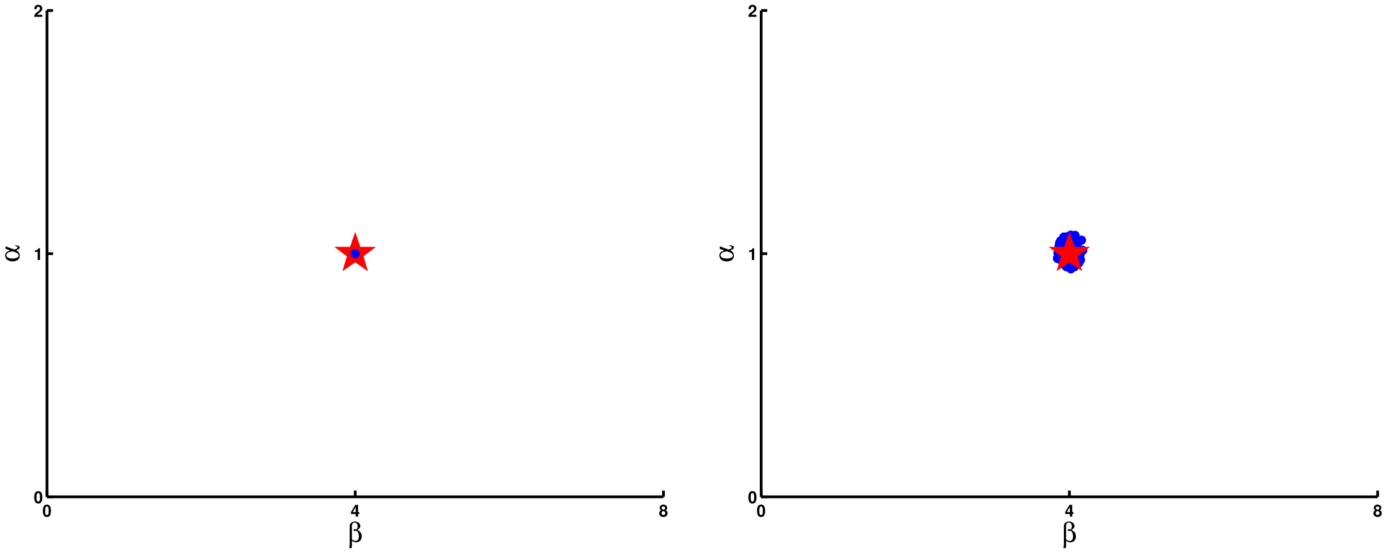
are less than 6%. We then conclude that both parameters, β and α, are practically identifiable given prevalence data in the basic SIR model. Our findings here agree with Capaldi's results [4], even though we implement a different noise structure, simulate different number of data points and employ a different numerical algorithm.

Fig. 1 is the scatter plot of the parameters' estimates from the generated data sets with $[\widehat{\beta}, \widehat{\alpha}] = [4, 1]$ when $\sigma_0 = 0\%$ and 10%. The scatter plots are generated by taking the true parameter value as the center of the plot, that is the β axis ranges from [0, 8] and α axis ranges from [0, 2]. As expected from Table 3, when there is no error, both parameters are exactly estimated. When the noise level goes up to 10%, there is a small variation among the 1000 estimates of the parameters. Nevertheless, they are still close to their true values, which confirms the above conclusion.

Fig. 2 provides an informative look at the effect of R_0 on the correlation coefficient $\chi = \chi_{\alpha\beta}$ as given in (4.4) and the parameters' variation. Here, we consider 3 levels of transmissibility, using R_0 values of 2, 4, and 8, respectively. In each case, we take α to be 1, corresponding with the infectious period of 1 time unit. Accordingly, because $R_0 = \dfrac{\beta}{\alpha}$, we take β to be 2, 4, and 8. For smaller R_0, the contours of the cost function are approximated by ellipses whose major axes are oriented along the line $\beta = R_0\alpha$, indicating the strong correlation between the two estimates of β and α. For values of R_0 that lead to lower (but still positive) correlation, the contours are closer to being circular. However, after R_0 reaches a certain threshold, the variation of β estimate starts to increase. The large variation in the estimate of β is shown in Fig. 2 where the contours are approximated by long thin horizontal ellipses. The difficulty while estimating β when R_0 is too large is the result of transmission events occurring dominantly at the beginning of timespan, yielding fewer informative data points [4].

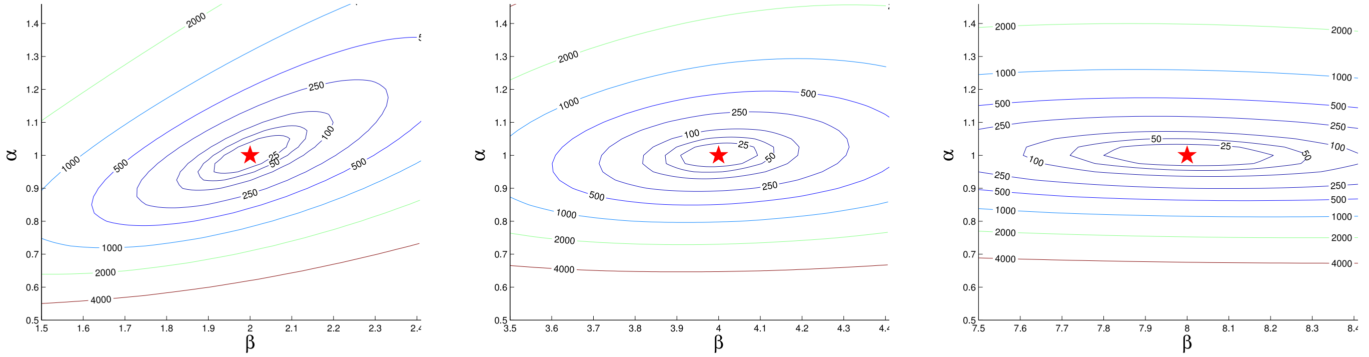
Finally, calculating the Fishers Information Matrix as given in (4.4), we compute the correlation coefficient $\chi_{\alpha\beta}$. Because the σ factors cancel out in the correlation coefficient formula, at all error levels, $\chi_{\alpha\beta} = 0.211$. We also calculate the correlation coefficient $\rho_{\alpha,\,\beta}$ given in (4.2) using the estimates from the Monte Carlo simulation. Averaging these values across all error levels, we obtain $\overline{\rho}_{\alpha\beta} = -0.092$. In both calculations, the magnitudes of this correlation stay close to 0, indicating that the two parameters β and α are distinguishable.

In summary, SIR model is both structurally and practically identifiable from prevalence data. Next, we investigate the evolution of the identifiability. In the case of an emerging infection, the parameters of the model are estimated as the data provided. Hence, the estimates are done at the initial stage of the outbreak where the number of infected individuals exhibits an exponential rise. Investigating the model's identifiability from partial time-course data will help us assess the reliability of the parameter estimation updates and potentially improve the projection the currently-evolving epidemics. We further examine



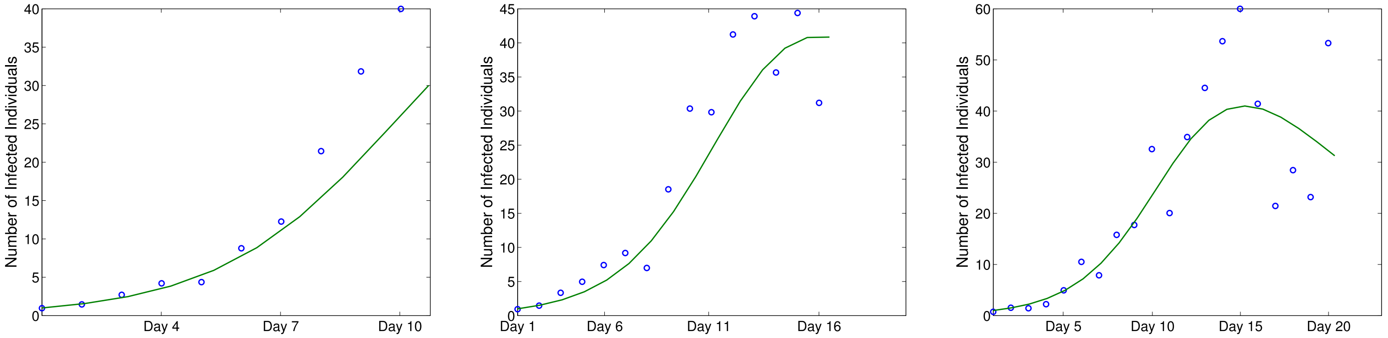
(A) $\sigma_0 = 0\%$ (B) $\sigma_0 = 10\%$

Fig. 1. Monte Carlo simulation results: Scatter plots of the parameters of the SIR Model (2.1) from 1000 fitting to prevalence data. The true parameters $[\beta, \alpha] = [4,1]$ are shown as red stars. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



(A) $[\beta,\alpha]=[2,1], \chi=0.736$ (B) $[\beta,\alpha]=[4,1], \chi=0.211$ (C) $[\beta,\alpha]=[8,1], \chi=-0.132$

Fig. 2. Contours of the error function in (3.5) given in the (β, α)-plane for the SIR Model (2.1) for prevalence data.



(A) 10 data points (B) 15 data points (C) 20 data points

Fig. 3. Partial time course of prevalence data from the SIR model.

Table 4
Monte Carlo simulation results: ARE for each parameter of the SIR model (2.1) when fitted to prevalence data of partial time course. Subscripts indicate number of data points. Initial guess for the iterative numerical optimization algorithm is $[\beta_0, \alpha_0] = [0.1, 0.1]$.

Error level (%)	β_{10}(%)	α_{10}(%)	β_{15}(%)	α_{15}(%)	β_{20}(%)	α_{20}(%)
0	5.17E−07	1.59E−06	1.29E−08	4.07E−07	1.56E−07	2.52E−07
1	1.37	4.38	0.38	0.92	0.17	0.35
5	7.06	22.78	1.95	4.76	0.88	1.85
10	13.74	44.1	3.65	9.10	1.69	3.53
20	23.74	76.00	8.00	19.97	3.70	7.13
30	28.6	91.56	11.74	29.25	5.32	11.33

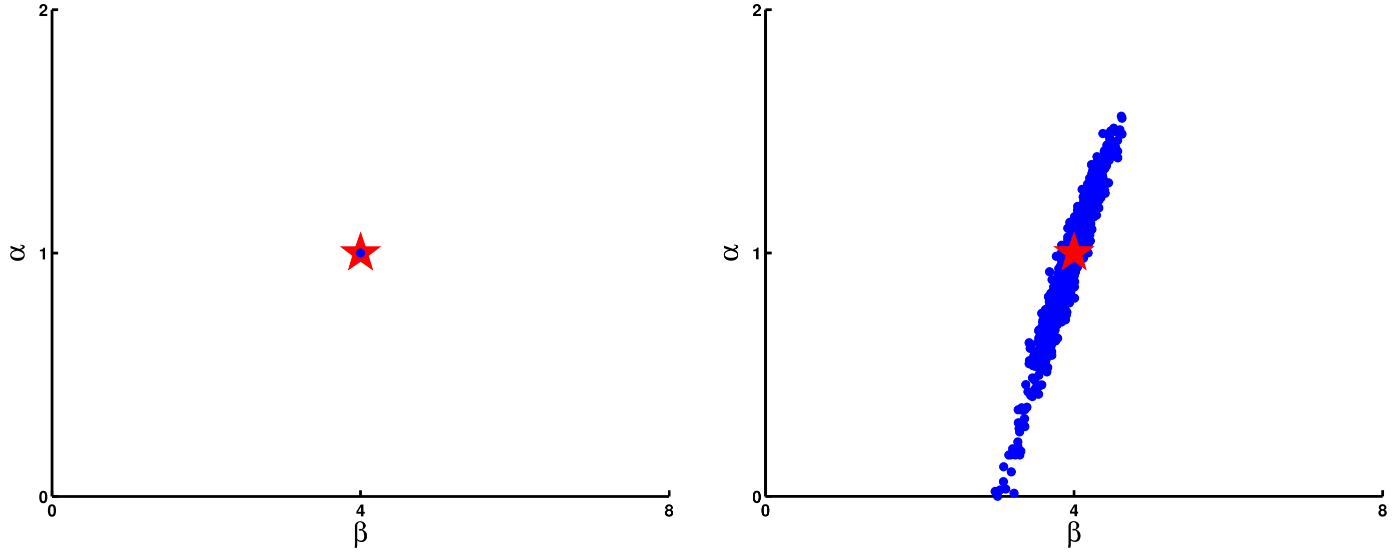
Table 5
Monte Carlo simulation results: ARE for each parameter of the SIR model (2.1) when fitted to cumulative incidence data.

Error level (%)	β(%)	α(%)
0	1.59E−07	3.68E−07
1	0.501	1.920
5	2.552	9.478
10	5.365	20.631
20	10.738	41.769
30	14.891	57.647

the evolution of practical identifiability of the SIR model as the prevalence data arrive in real time. Instead of a full time course with 51 data points, we consider (i) 10 data points that only capture the period of increasing number of cases, (ii) 15 data points that just reach the peak of prevalence, and (iii) 20 data points that end shortly after the peak of prevalence (Fig. 3). In data sets with 10 observations, as the error level increases, the relative errors of both parameters β_{10} and α_{10} increase to larger than the implemented error. Thus, we conclude that, given prevalence data available only exponential rise of the infected individuals, the SIR model, β and α are not identifiable. In data sets with at least 15 observations, even though the ARE of the parameters β_{15}, β_{20}, α_{15} and α_{20} are larger than those in Table 4, these values remain smaller than the error levels. We conclude that the SIR model becomes practically identifiable to prevalence data when the outbreak reaches its peak.

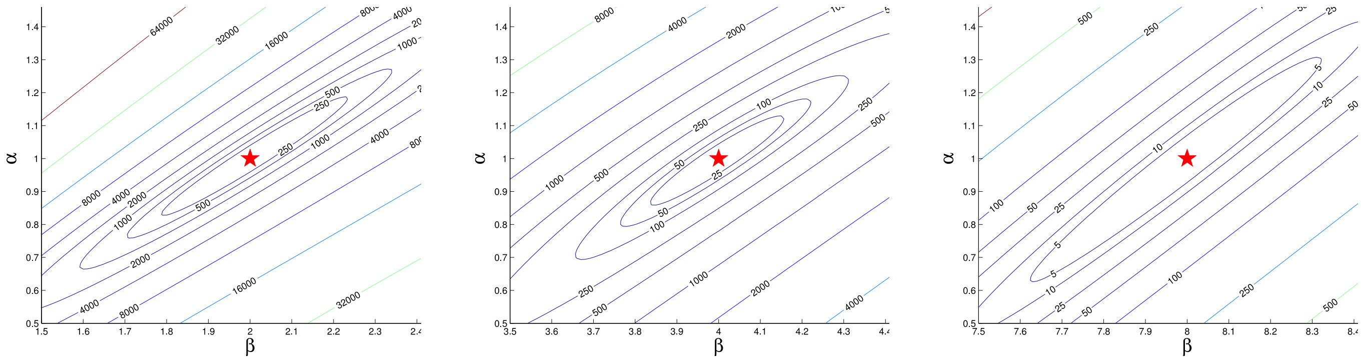
5.1.2. Experiment B: cumulative incidence data

The SIR model (2.1) is structurally identifiable from the cumulative incidence observations. To test the practical identifiability of the model parameters, we first perform Monte Carlo simulations by taking $[\hat{\beta}, \hat{\alpha}] = [4, 1]$ as the true parameter set. We start the iterative algorithm Nelder–Mead with initial guess of $[\beta_0, \alpha_0] = [0.1, 0.1]$. The AREs of the parameters β and α at different measurement error levels are presented in Table 5. When there is no measurement error ($\sigma_0 = 0\%$), both parameters are well identified (max ARE is of order 10^{-7}), which confirms the theoretical identifiability result. As expected, as the error level increases, both parameters' AREs increase, although much quicker than



(A) $\sigma_0 = 0\%$ (B) $\sigma_0 = 10\%$

Fig. 4. Monte Carlo simulation results: Scatter plots of the parameters of the SIR Model (2.1) from 1000 fitting to cumulative incidence data. The true parameters [β, α] = [4,1] are shown as red stars. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



(A) $[\beta, \alpha] = [2, 1], \chi = 0.991$ (B) $[\beta, \alpha] = [4, 1], \chi = 0.943$ (C) $[\beta, \alpha] = [8, 1], \chi = 0.977$

Fig. 5. Contours of the error function in (3.5) given in the (β, α)-plane for the SIR Model (2.1) for the cumulative incidence data.

Table 6
Monte Carlo simulation results: ARE for each parameter of the SEIR model (2.2) when fitted to prevalence data. Initial guess for the iterative numerical optimization algorithm is [β_0, α_0, η_0] = [0.1, 0.1, 0.1].

Error level (%)	β(%)	α(%)	η(%)
0	2.81E − 07	5.50E − 08	4.03E − 07
1	1.16	0.15	1.93
5	6.33	0.76	9.89
10	11.49	1.52	18.02
20	18.37	3.14	252.14
30	21.47	4.43	827.48

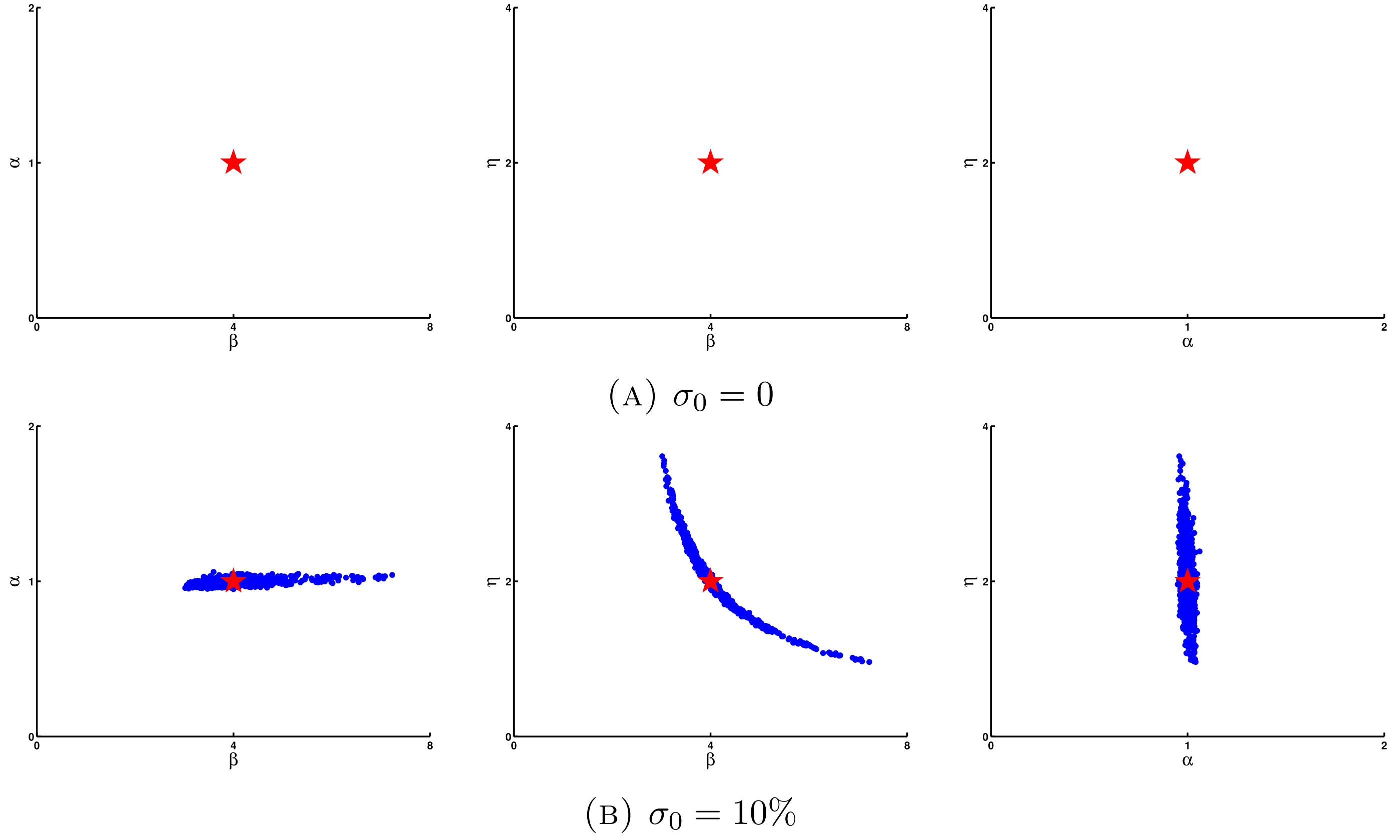
Table 7
Monte Carlo simulation results: ARE for each parameter of the SEIR model (2.2) when fitted to prevalence data. Initial guess for the iterative numerical optimization algorithm is [β_0, α_0, η_0] = [4, 1, 2].

Error level (%)	β(%)	α(%)	η(%)
0	0	0	2.22E − 14
1	1.12	0.15	1.84
5	5.94	0.75	9.38
10	12.09	1.49	17.91
20	16.97	2.91	30.39
30	22.56	4.36	47.66

the prevalence case, especially the ARE of parameter α. When the error level is at 20%, the value of α's ARE reaches over 40%, whereas in the prevalence case, it is less than 4%. Since the ARE of the parameter α is consistently higher than the measurement error level, we conclude that the parameter α is practically unidentifiable in the SIR model when the cumulative incidence data are considered.

Fig. 4 is the scatter plot of the parameters' estimates from the generated data sets with true parameter values [β, α] = [4, 1] when σ_0 = 0% and 10%. As observed from Table 5, when there is no error, both parameters are exactly estimated. When the noise level goes up to 10%, there is a linear trend in the estimates. This suggest a high correlation between the parameters, which is explained in more depth with the contour plots in Fig. 5.

Fig. 5 shows the contours of the cost function on the (β, α) plane given the cumulative incidence data. To be consistent with the prevalence case, in this practical identifiability of the SIR model with cumulative incidence data, we also consider 3 levels of transmissibility, using $\mathscr{R}_0$ values of 2, 4, and 8, respectively. Similarly, in each case, we take α to be 1 and β to be 2, 4, and 8. In all three cases, the general shape of the contours are very similar: very long thin ellipses of which



(A) $\sigma_0 = 0$

(B) $\sigma_0 = 10\%$

Fig. 6. Monte Carlo simulation results: Scatter plots of the parameters of the SEIR Model (2.2) when fitted prevalence data. The true parameters $[\hat{\beta}, \hat{\alpha}, \hat{\eta}] = [4, 1, 2]$ are shown as red stars.

major axes are oriented along the line $\beta = R_0\alpha$. This implies that there is no strong effect of the magnitude of R_0 on the correlation between the estimates of the parameters. Specifically, the two parameters are highly correlated at any R_0 value.

Finally, from the Fishers Information Matrix, we compute the correlation coefficient between α and β. At all error levels, $\chi_{\alpha\beta} = 0.864$. Averaging these values across all error levels using the estimates from the Monte Carlo simulation, we obtain $\bar{\rho}_{\alpha\beta} = 0.964$. Although there is a slight difference between the values computed using the correlation matrix and the Monte Carlo simulation, the magnitudes of the correlation are very close to 1, implying a high correlation between the two parameters β and α, which verifies what we observe in Figs. 4 and 5. Based on the numerical experiments, we conclude that even though the SIR model is structurally identifiable from the cumulative incidences observations, practically it is unidentifiable.

5.2. SEIR Model

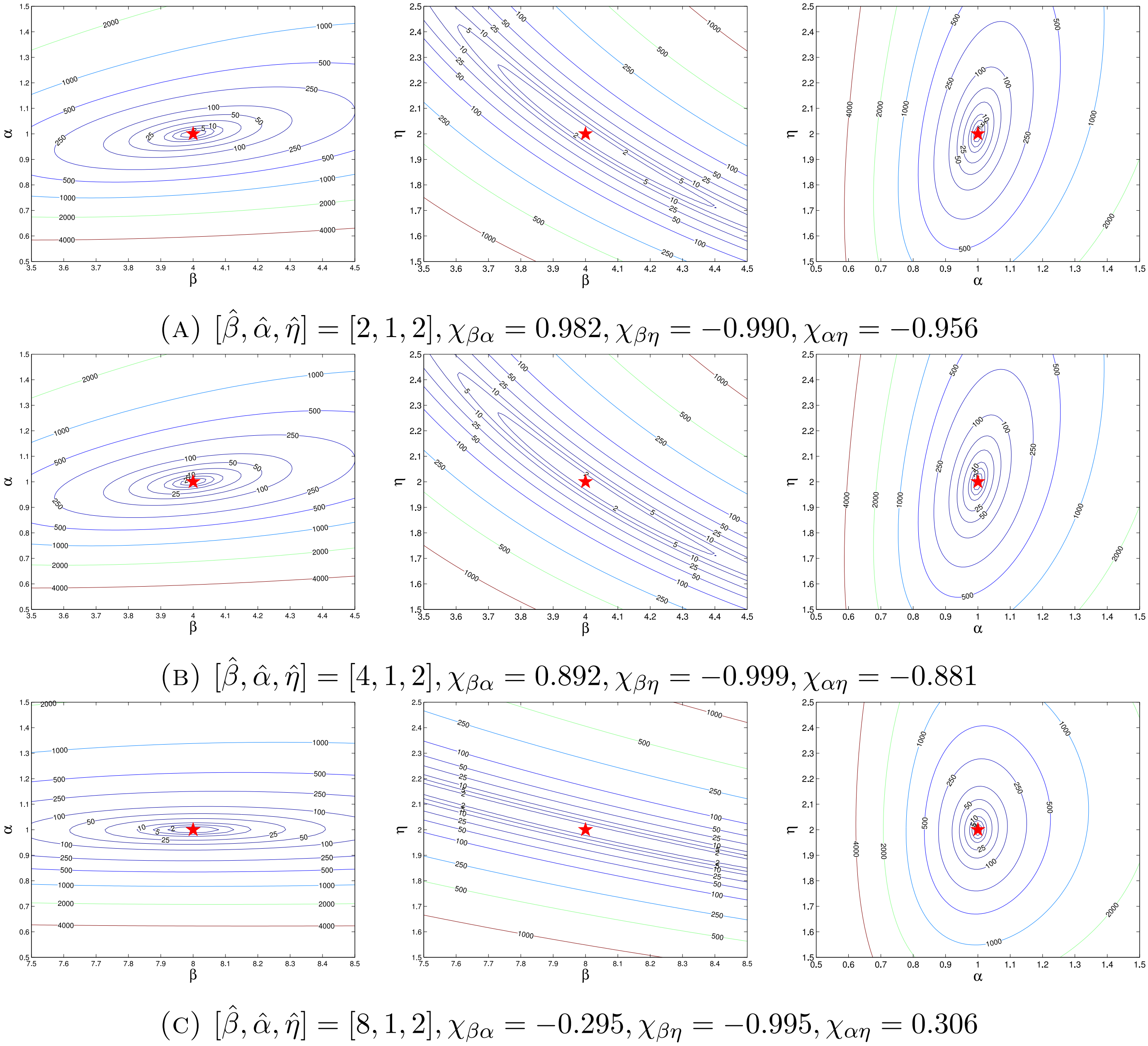
5.2.1. Experiment A: prevalence data

differential algebra analysis of the SEIR model (2.2) states that the model is locally structurally identifiable from observations of the prevalence. We continue our analysis with practical identifiability of the SEIR model. Similar to the previous case of the SIR model, we generate 1000 replicates of data at 51 time points according to the Monte Carlo simulation procedure. Also using `fminsearchbnd`, we fit the model to each of the 1000 replicate data sets with the starting guess of $\beta_0 = \alpha_0 = \eta_0 = 0.1$. The AREs at six measurement error levels, $\sigma_0 = 0, 1, 5, 10, 20$, and 30% are shown in Table 6, given the true parameters $[\hat{\beta}, \hat{\alpha}, \hat{\eta}] = [4, 1, 2]$.

When there is no measurement error ($\sigma_0 = 0\%$), all three parameters are well identified (max ARE is of order 10^{-7}). As the error level increases up to 5%, the relative errors of α increases but remains well below the error level, while the relative errors of β and η, becomes larger than the error level. Particularly, after the implemented error reaches 20%, ARE of the rate of becoming infections parameter η increases tremendously. Hence, we shall derive that α is practically identifiable given prevalence data in the SEIR model, while β and η are not. To rule out the possible source of unidentifiability due to model being only locally structurally identifiable, in the next experiment we start the numerical optimization at the true value. In Table 7, we note that similar ARE values are obtained even when we start with our parameter guesses at the true values: $[\beta_0, \alpha_0, \eta_0] = [\hat{\beta}, \hat{\alpha}, \hat{\eta}] = [4, 1, 2]$. Thus, despite being structurally locally identifiable, practically, this model is not identifiable to prevalence data.

Fig. 6 is the scatter plot of the parameters' estimates from the generated data sets with $[\hat{\beta}, \hat{\alpha}, \hat{\eta}] = [4, 1, 2]$ when the level error $\sigma_0 = 0\%$ and 10%. As we expected from Table 6, when there is no error, all parameters are exactly estimated. When the noise level goes up to 10%, while the estimates for β and α stay relatively close to their true values, the estimates for η have a much larger variation.

Additionally, we can infer the effect of $\mathscr{R}_0$ on the correlation coefficient $\chi = \chi_{\alpha\beta}$ as given in (4.4) and the parameters' variation from Figs. 7 and 8. Consistent to the SIR model, in the contour plot of the error function of Fig. 7, we again consider 3 levels of transmissibility: $\mathscr{R}_0 = 2, 4$, and 8. Recalling that for the SEIR model, $\mathscr{R}_0 = \dfrac{\beta}{\alpha}$, we put η equal 2, α equal 1 and β equal 2, 4, and 8, respectively. For smaller $\mathscr{R}_0$, the contours of the cost function on the (α, β)-plane are approximated by rounder curves centered at the true values. On this same plane, as $\mathscr{R}_0$ increases, these contours become flatter, horizontal ellipses, indicating the much larger variation increase in β estimates relative to α. Furthermore, on the (η, β), the contours are represented as long, thin, downward-oriented ellipses, indicating large correlation values between β and η at all $\mathscr{R}_0$ levels. Finally, the correlation between α and η seems to decrease as $\mathscr{R}_0$ increases. We can also inspect the dependence



(A) $[\hat{\beta}, \hat{\alpha}, \hat{\eta}] = [2, 1, 2], \chi_{\beta\alpha} = 0.982, \chi_{\beta\eta} = -0.990, \chi_{\alpha\eta} = -0.956$

(B) $[\hat{\beta}, \hat{\alpha}, \hat{\eta}] = [4, 1, 2], \chi_{\beta\alpha} = 0.892, \chi_{\beta\eta} = -0.999, \chi_{\alpha\eta} = -0.881$

(C) $[\hat{\beta}, \hat{\alpha}, \hat{\eta}] = [8, 1, 2], \chi_{\beta\alpha} = -0.295, \chi_{\beta\eta} = -0.995, \chi_{\alpha\eta} = 0.306$

Fig. 7. Contours of the error function in (3.5) given in the (β, α), (η, β), and (η, α)-planes for SEIR Model (2.2), fitted to prevalence data.

of the correlation coefficient and standard errors for estimates from Fig. 8. When $\mathscr{R}_0$ is too small (less than a threshold of approximately 1.8), there is large variation within each parameter's estimates (Fig. 8a). When $\mathscr{R}_0$ goes beyond this threshold, the standard errors stay at a reasonable magnitude, but β and η are virtually indistinguishable due to their almost perfect correlation: $\chi_{\beta\eta} \approx -1$ (Fig. 8b). Thus, there is not a value of $\mathscr{R}_0$ where the estimates are stable as well as all the parameters are identifiable.

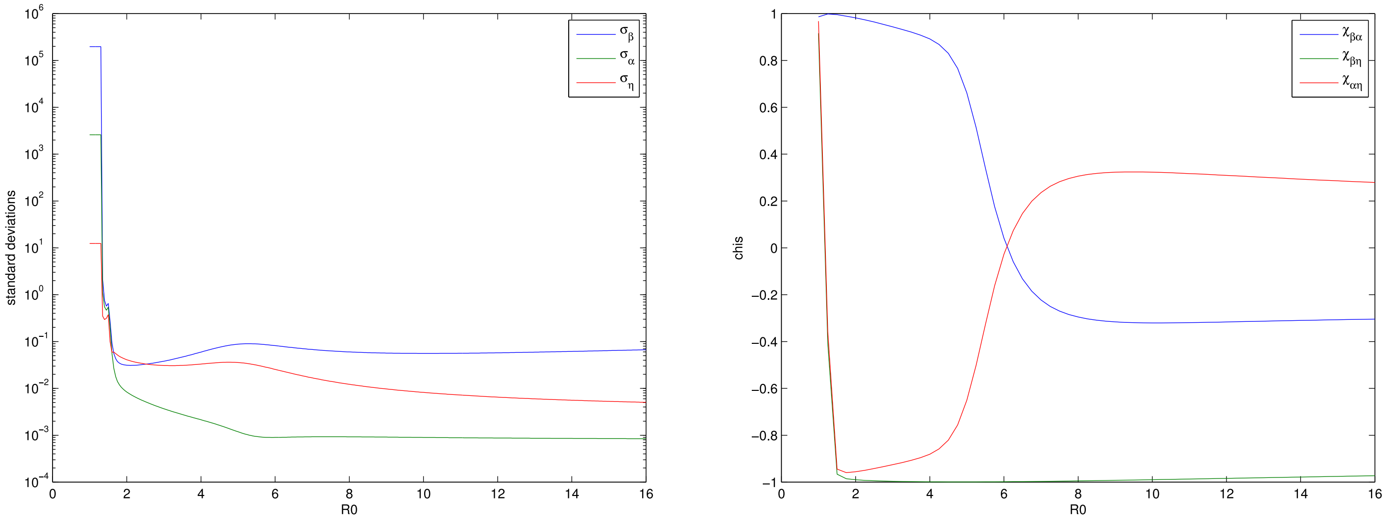
5.2.2. *Experiment B: cumulative data*

Performing the Monte Carlo simulations, we calculate the AREs at six measurement error levels, and scatter plot the estimates of the parameters β, α, and η. This time, we start with our initial guesses at the true values: $[\beta_0, \alpha_0, \eta_0] = [\hat{\beta}, \hat{\alpha}, \hat{\eta}] = [4, 1, 2]$. The results are shown in Table 8 and Fig. 9.

When there is no measurement error ($\sigma_0 = 0\%$), all three parameters are well identified (max ARE is of order 10^{-6}). As the error level increases up to 10%, the relative errors of all three parameters increase rapidly and stay much larger than the implemented error. Our numerical algorithm reported a runtime error when the noise level exceed 10% and couldn't find proper estimates of the parameters. Thus, we infer that, given cumulative incidence data in the SEIR model, β, α and η are not identifiable. The following scatter plot in Fig. 9 helps us better visualize the degrees of variation between the parameters' estimates. When σ_0 goes from 1% to 5%, the estimates spread out remarkably. At $\sigma_0 = 10\%$, the substantial variation in the estimates of α is seen from its broad range in the scatter plot.

From Fig. 10, we can observe that at all $\mathscr{R}_0$ values, the α estimates have high variation, indicated by the parallel contour lines. This is consistent with our findings in Table 8, where α's ARE are very large when error level is just above 0%. Moreover, when $\mathscr{R}_0$ is small, there is a high correlation between β and η, which is implied by the very long, thin ellipses. As $\mathscr{R}_0$ rises, the magnitude of this correlation seems to decrease.

Fig. 11 shows the dependence of the correlation coefficient and standard errors for estimates in the SEIR model with cumulative



(A) Estimates' standard errors on a log scale

(B) Correlation coefficients χ's

Fig. 8. Dependence of the correlation coefficient and standard errors for estimates of β, α, and η on the value of $\mathscr{R}_0$, SEIR model, prevalence data.

Table 8
Monte Carlo simulation results: ARE for each parameter of SEIR model (2.2) when fitted cumulative incidence data. Initial guess for the iterative numerical optimization algorithm is the true parameter $[\beta_0, \alpha_0, \eta_0] = [4, 1, 2]$.

Error level (%)	β(%)	α(%)	η(%)
0	0	0	2.22E − 14
1	8.39	9.23	7.24
5	33.80	31.96	32.28
10	44.62	46.11	40303.44

incidence data. Although the standard errors of each parameter's estimates stay quite reasonable (Fig. 11a), there is not a value of $\mathscr{R}_0$ where all three correlation values are simultaneously less than 0.5 in magnitude (Fig. 11b). Therefore, these parameters are not identifiable.

5.3. Modifying the SEIR model

5.3.1. Prevalence data: SEIR model with fixed η

In Table 6 of the above prevalence case, when $\sigma_0 > 0$, we observed the practical unidentifiability of η, the rate of becoming infectious parameter. To modify this model, we now assume that η is calculated prior to the numerical estimation with `fminsearch`. Once again, we carry out the Monte Carlo simulations with the objective of estimating β and α, using the starting guess of $\beta_0 = \alpha_0 = 0.5$. The ARE's and scatter plots of these two parameters' estimates are presented in Table 9 and Fig. 12. After we fix the value of η, the AREs of both β and α stay well under the error level σ_0 at all values of σ_0, meaning that the parameters of the SEIR model with prior knowledge of η is practically identifiable from the prevalence data.

5.3.2. Cumulative incidence data: SEIR model with fixed α or η.

From the preceding analysis on SEIR model with cumulative incidence data, we know that all three parameters β, α, and η, are practically unidentifiable. However, as demonstrated in Table 8, when σ_0 reaches 10%, ARE$_\alpha$ jumps immediately to over 40,000%, extremely more rapidly than ARE$_\eta$ and ARE$_\beta$. Therefore, in this case, hoping to make the model identifiable, we fix the value for α before carrying out the Monte Carlo simulations. Putting $\hat{\alpha} = 1$, we estimate β and η using `fminsearch` with the starting guess of $\beta_0 = 4$, $\eta_0 = 2$ (true values.) From Table 10, Figs. 13 and 10(B) we see that even when we fix α and set the initial guesses at the true parameters' values, assuming a positive noise level, the two remaining parameters, β and η are still highly correlated and unidentifiable.

Next, we fix η to its true value, and perform the Monte Carlo simulations. As the results show in Table 11, parameters in this modified SEIR model are practically unidentifiable when fitted to cumulative incidence data. The high AREs are due to high correlations between parameters (see Fig. 10(B)).

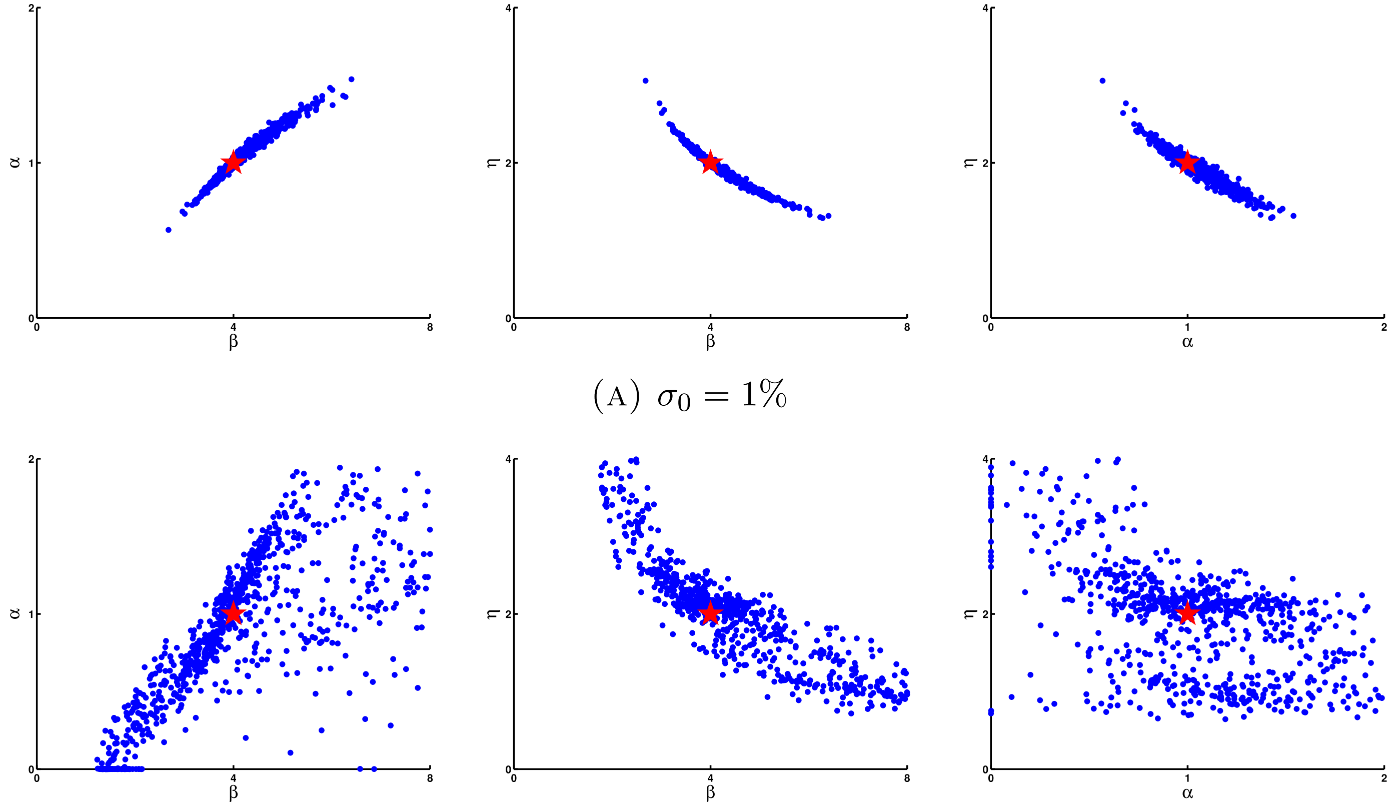
5.4. Fitting SEIR model to Ebola outbreak in West Africa in 2014

We fit the SEIR model on the reported cumulative number of cases of the 2014 Ebola outbreak in Liberia. This cumulative incidence data set was obtained from the WHO website with 22 data points recorded for over two months, from June 16, to August 20, 2014. Following Althaus [1], we assume the total population size $N = S + E + I + R = 10^6$ individuals and fix $\eta = 0.189$, using the `fminsearchbnd` we estimate the parameters as $\beta = 0.1435$ and $\alpha = 0.0718$. The model output with the estimated values are plotted together with the WHO cumulative incidence Ebola data in Fig. 14. We then use the fitted parameters $\beta = 0.1435$ and $\alpha = 0.0718$ as true parameter values in the succeeding Monte Carlo simulation, where we set initial guesses as the fitted parameter values. Consistent with what we found in simulated data (Table 11), Table 12 shows the SEIR model parameters are practically unidentifiable when fitted to the real cumulative incidence Ebola data.

5.5. Treatment model

5.5.1. Experiment A: treatment data

We know that the treatment model (2.3) is not structured to identify its parameters from the observations of the treatment state variable. Hence, it would be practically unidentifiable as well. In fact, we do not need to perform practical identifiability analysis for the treatment model (2.3). However, for some epidemic models structural identifiability analysis might not be available. Note that differential algebra approach requires the dynamical system (3.1) to be a rational function of state variables. If structural identifiability analysis is not possible for the epidemic model, then practical identifiability analysis would be the only way to study the identifiability of the model parameters. Thus, we solely perform the Monte Carlo simulations for the treatment model by generating 100 replicates of data 51 time points Also using



(A) $\sigma_0 = 1\%$

(B) $\sigma_0 = 10\%$

Fig. 9. Monte Carlo simulation results: Scatter plots of the parameters of the SEIR Model (2.2) when fitted to 1000 replicated of cumulative incidence data. The true parameters $[\hat{\beta}, \hat{\alpha}, \hat{\eta}] = [4, 1, 2]$ are shown as red stars. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

`fminsearchbnd`, we fit the model to each of the 100 replicate data sets with the initial guesses at the true values: $[\beta_0, \alpha_0, \delta_0, \gamma_0, \nu_0] = [\hat{\beta}, \hat{\alpha}, \hat{\eta}, \hat{\gamma}, \hat{\nu}] = [4, 1, 0.8, 5, 2]$. Table 13 displays the ARE values for the five parameters β, α, δ, γ and ν at six different noise values of σ_0. It shows that all five parameters are unidentifiable. While the relative error of β, γ, and ν seems to remain under 100%, relative error of the other two parameters goes up more drastically. To be specific, ARE_α shoots up to over 68% even when σ_0 just reaches 5%.

5.5.2. *Experiment B: cumulative incidence data*

The treatment model (2.3) is not structured to identify its parameters from the cumulative incidences data, hence no need to perform any practical identifiability analysis. We just carry out the same Monte Carlo simulation procedure as in the treatment case. Also using `fminsearchbnd`, we fit the model to each of the 100 replicate data sets with the same initial guesses at the true values: $[\beta_0, \alpha_0, \delta_0, \gamma_0, \nu_0] = [\hat{\beta}, \hat{\alpha}, \hat{\eta}, \hat{\gamma}, \hat{\nu}] = [4, 1, 0.8, 5, 2]$. Once again, by displaying the ARE values for the five parameters β, α, δ, γ and ν at six different noise values, Table 14 shows that all five parameters are unidentifiable; and once again, we observe that the increase in the relative errors of α and δ is a lot more rapid. Specifically, ARE_α shoots up to almost 500% even when σ_0 just reaches 5%.

5.6. *Modified treatment model*

5.6.1. *Treatment data: fixed δ and γ*

Recall in Remark 3, we noted that if the parameters δ and γ are fixed, the treatment model becomes structurally identifiable. In order to discover if this identifiability holds in practice, we proceed with Monte Carlo simulation after fixing δ and γ at 0.8 and 5, respectively. We now estimate β, α and ν using `fminsearch` with the starting guess of $\beta_0 = 4$, $\alpha_0 = 1$ and $\nu_0 = 2$ (true values). In Table 15, we observe that the relative errors of all parameters are always larger than the implemented noise. Therefore, we conclude that fixing δ and γ before estimating the other parameters would still result in unidentifiability of the remaining parameters.

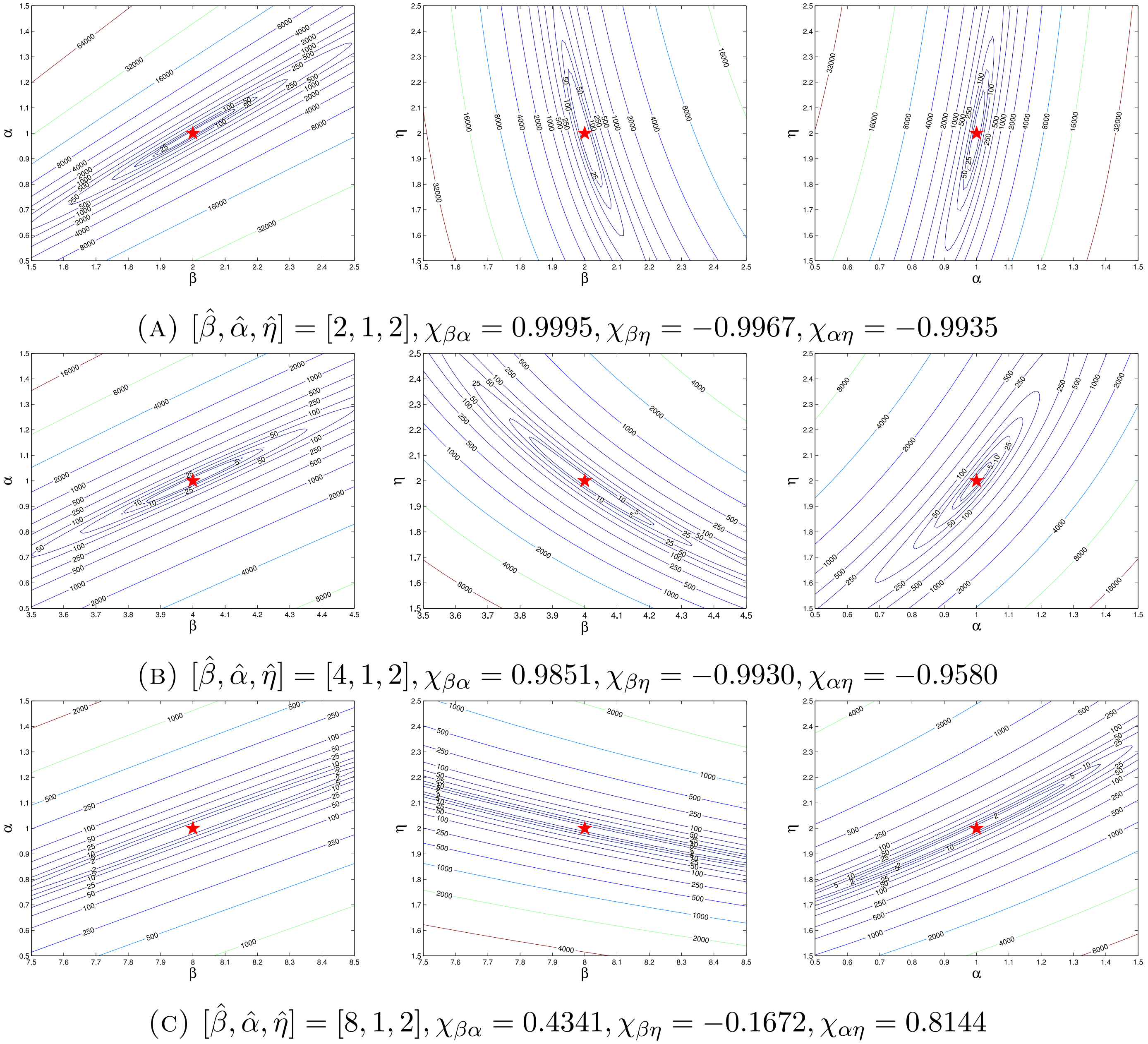
Since α appears to be the troublesome parameter, we go a step further and fix this parameter, rerun the numerical method with starting guess of $[\beta_0, \nu_0] = [0.1, 0.1]$. As the ARE result shows in Table 16, the remaining parameters, β and ν, are identifiable.

5.6.2. *Treatment data: fixed α and δ*

In Table 13, we observed α and δ to be the two most troublesome parameters with their rapid rate of increase in ARE. We want to see if fixing these two parameters makes the model identifiable, and thus execute the Monte Carlo simulation after fixing $\alpha = 1$ and $\delta = 0.8$. We estimate the remaining parameters β, γ and ν using `fminsearch` with the starting guess of $\beta_0 = 4$, $\gamma_0 = 5$ and $\nu_0 = 2$ (true values). In Table 17, we observe, while the relative errors of β and ν remains small, that of γ increases faster than the added noise. Therefore, we conclude that fixing α and δ before estimating the other parameters still does not resolve the unidentifiability issue of the remaining parameters.

5.6.3. *Cumulative incidence data: fixed α and δ*

Recall in Remark 2 we noted that if the parameters α and δ are fixed, the treatment model becomes structurally identifiable. In order to discover if this identifiability holds in practice, we proceed with Monte Carlo simulation after fixing α and δ at 0.8 and 1, respectively. We now estimate β, α and ν using `fminsearch` with the starting guess of $\beta_0 = 4$, $\gamma_0 = 5$ and $\nu_0 = 2$ (true values). In Table 18, we observe that while the relative error of β increases reasonably as σ_0 increases, that of γ and ν is always larger than the implemented noise. Therefore, we



(A) $[\hat{\beta}, \hat{\alpha}, \hat{\eta}] = [2, 1, 2], \chi_{\beta\alpha} = 0.9995, \chi_{\beta\eta} = -0.9967, \chi_{\alpha\eta} = -0.9935$

(B) $[\hat{\beta}, \hat{\alpha}, \hat{\eta}] = [4, 1, 2], \chi_{\beta\alpha} = 0.9851, \chi_{\beta\eta} = -0.9930, \chi_{\alpha\eta} = -0.9580$

(C) $[\hat{\beta}, \hat{\alpha}, \hat{\eta}] = [8, 1, 2], \chi_{\beta\alpha} = 0.4341, \chi_{\beta\eta} = -0.1672, \chi_{\alpha\eta} = 0.8144$

Fig. 10. Contours of the error function in (3.5) for the SEIR Model (2.2) when fitted cumulative incidence data.

conclude that fixing δ and α before numerically estimating the other parameters would still result in unidentifiability of the γ and ν.

In an attempt to obtain a practically identifiable parameters, we further fix the parameter γ and perform the Monte Carlo simulations. As we see in Table 19, ν remains unidentifiable, indicating the only identifiable parameter is β. When we go one step further and fix ν, the model becomes identifiable with its only parameter β identified.

5.6.4. Treatment and cumulative incidence data

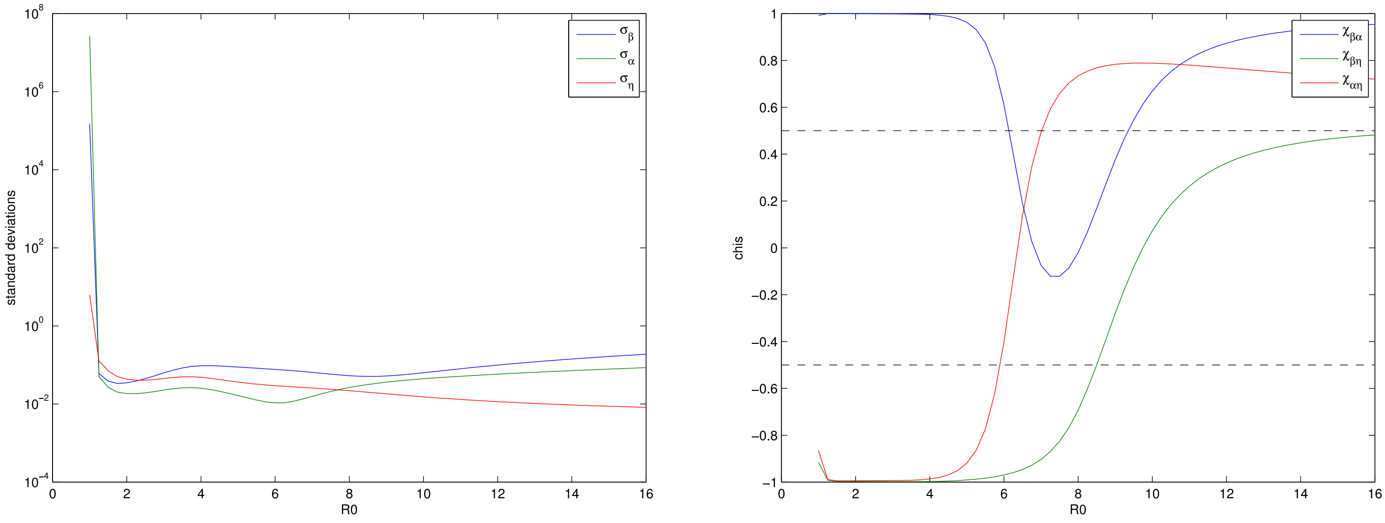
Recall in Remark 3 we noted that if we have both treatment and cumulative incidence data, then the treatment model (2.3) becomes locally identiable. Furthermore, if ν is fixed, the model becomes structurally identifiable. In order to analyze this identifiability in practice, we proceed with Monte Carlo simulation assuming both treatment and cumulative incidence data observed, before and after fixing ν. In both cases, we estimate all parameters starting at the true values. In Table 20, we observe that the relative errors of all parameters are always larger than the implemented noise. Therefore, we conclude, even when both treatment and cumulative incidence data are considered, the parameters remain unidentifiable.

After we fix ν on the observed treatment and cumulative incidence data, the parameters' relative errors are displayed in Table 21. We observe that the relative errors of all parameters are always larger than the implemented noise. Thus, we deduce the model's unidentifiability (only the parameter β is practically identifiable), even when both treatment and cumulative incidence data are considered and ν is fixed before fitting other parameters.

Once again, attempting to acquire an identifiable model, we perform Monte Carlo simulations when fitted to both data sets with ν and δ fixed. Nevertheless, as seen from the AREs of the parameters in Table 22, only β is practically identifiable.

6. Conclusion

Mathematical models are used as a tool to combat an emerging infectious disease by projecting potential cases, estimating key



(A) Estimates' standard errors on a log scale (B) Correlation coefficient ρ

Fig. 11. Dependence of the correlation coefficient and standard errors for estimates of β, α, and η on the value of $\mathscr{R}_0$, SEIR model, cumulative incidence data.

Table 9
Monte Carlo simulation results: ARE for each parameter of the SEIR model (2.2) when fitted to prevalence data with fixed η. Initial guess for the iterative numerical optimization algorithm is $[\beta_0, \alpha_0] = [0.5, 0.5]$.

Error level (%)	β(%)	α(%)
0	1.74E−07	4.26E−07
1	0.09	0.16
5	0.45	0.82
10	0.93	1.62
20	1.88	3.38
30	2.82	5.21

Table 10
Monte Carlo simulation results: ARE for each parameter of the SEIR model (2.2) when fitted to cumulative incidences data with $\alpha = 1$ is fixed.

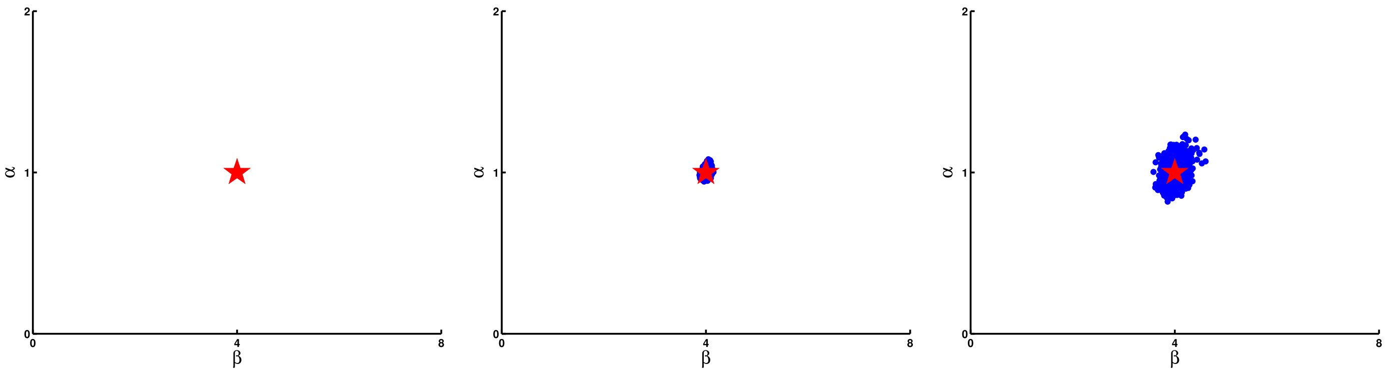
Error level (%)	β(%)	η(%)
0	0	2.22E−14
1	1.47	2.54
5	8.72	12.89
10	17.41	24.59
20	27.59	42.20
30	31.78	79.53

parameters of the outbreak, evaluating and optimizing control strategies. When a new disease emerges, understanding its characteristic parameters gives us more insight into the underlying mechanism of that particular disease, helps us timely predict its future epidemic, and perhaps even suggests the right control strategies to impede the diseases spread while optimizing its cost.

However, depending on the model and the observed data structure, the parameter estimation problem may not be well posed and yield inaccurate results. Two necessary conditions for the estimation problem to be well posed are the model's structural identifiability, which can be analyzed prior to collecting data, and its practical identifiability. While structural identifiability assumes ideal conditions such as error-free model and noise-free data, practical identifiability takes into account the noise in real-world data.

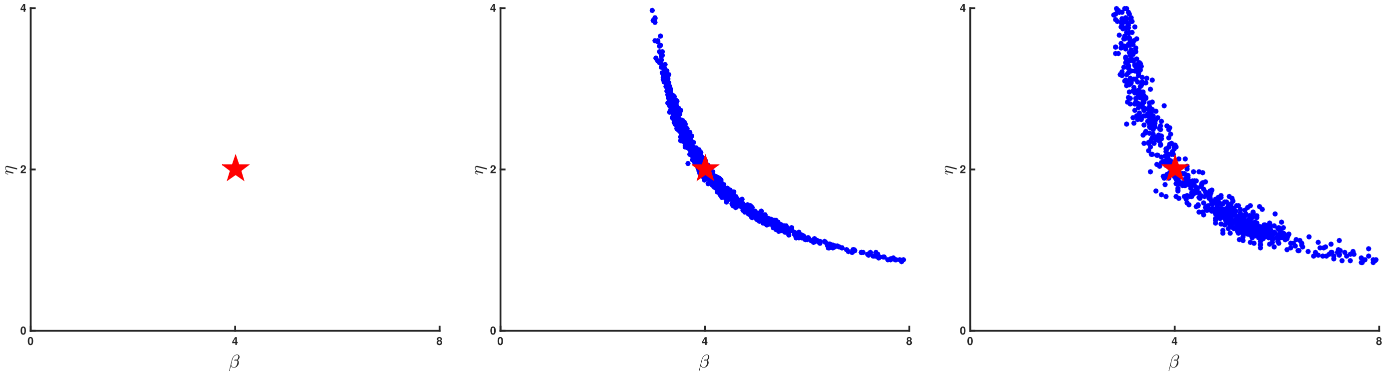
In this study, we analyzed both structural and practical identifiability of three basic outbreak models for an emerging disease: SIR, SEIR, and SITR. For SIR and SEIR models, we suppose we observe prevalence and cumulative incidences, for SITR we assumed that number of treated individuals are observed together with cumulative incidences.

We first used the differential algebra approach to determine each model's structural identifiability. From observations of the cumulative data, the parameters of the SIR and SEIR model can be structurally identified, but those of the treatment model cannot (Propositions 1, 3



(A) $\sigma_0 = 0$ (B) $\sigma_0 = 10\%$ (C) $\sigma_0 = 30\%$

Fig. 12. Monte Carlo simulation results: Scatter plots of the SEIR Model (2.2) when fitted to prevalence data with fixed $\eta = 2$. The true parameters $[\hat{\beta}, \hat{\alpha}] = [4, 1]$ are shown as red stars.



(A) $\sigma_0 = 0$ (B) $\sigma_0 = 10\%$ (C) $\sigma_0 = 30\%$

Fig. 13. Monte Carlo simulation results: Scatter plots of the parameters of the SEIR Model (2.2) when fitted to cumulative incidence data, with fixed $\alpha = 1$. The true parameters $[\hat{\beta}, \hat{\eta}] = [4, 2]$ are shown as red stars.

Table 11
Monte Carlo simulation results: ARE for each parameter of the SEIR model (2.2) when fitted to cumulative incidences data with $\eta = 2$ is fixed.

Error level (%)	β(%)	α(%)
0	0	0
1	1.11	2.75
5	5.45	13.40
10	11.45	28.45
20	22.01	54.83
30	33.14	81.39

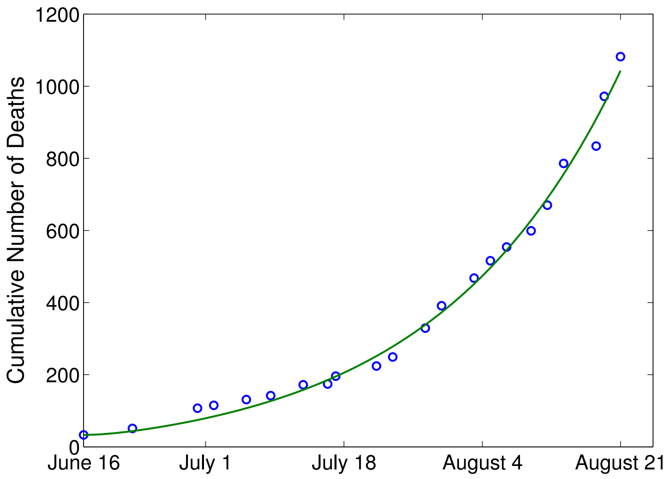


Fig. 14. Liberia's Ebola cumulative incidence data with fitted SEIR parameters: $\beta = 0.1435$ and $\alpha = 0.0718$.

Table 12
Monte Carlo simulation results: ARE for each parameter of the SEIR model (2.2) when fitted to real cumulative Ebola incidence data with fixed $\eta = 0.189$. Initial guess for the iterative numerical optimization algorithm is $[\beta_0, \alpha_0] = [0.1435, 0.0718]$.

Error level (%)	β(%)	α(%)
0	1.93E−14	1.93E−14
1	1.58	3.14
5	8.21	16.48
10	15.31	30.70
20	37.32	74.40
30	46.29	91.77

Table 13
Monte Carlo simulation results: ARE for each parameter of the Treatment model (2.3) when fitted to treatment data.

Error level (%)	β(%)	α(%)	δ(%)	γ(%)	ι(%)
0	0	0	1.39E−14	1.78E−14	2.22E−14
1	1.86	19.37	4.91	3.47	2.85
5	8.68	68.43	16.74	11.73	10.15
10	14.83	97.78	27.65	16.83	15.22
20	26.87	169.09	72.91	31.58	20.58
30	57.41	452.23	186.13	59.85	29.87

Table 14
Monte Carlo simulation results: ARE for each parameter of the Treatment model (2.3) when fitted to cumulative incidence data.

Error level (%)	β(%)	α(%)	δ(%)	γ(%)	ι(%)
0	0	0	1.39E−14	1.78E−14	2.22E−14
1	6.03	73.81	25.19	65.14	23.22
5	42.28	476.81	205.57	110.82	52.97
10	61.42	323.13	307.85	135.84	78.05
20	219.51	1185.79	1368.32	172.14	89.06
30	1410.10	11038.90	2239.73	563.07	115.11

Table 15
Monte Carlo simulation results: ARE for each parameter of the Treatment model (2.3) when fitted to treatment data with fixed δ and γ.

Error level (%)	β(%)	α(%)	ι(%)
0	0	0	2.22E−14
1	1.81	17.87	2.75
5	7.51	74.04	11.59
10	13.30	127.93	17.00
20	20.79	190.78	20.93
30	43.66	385.69	31.11

Table 16
Monte Carlo simulation results: ARE for each parameter of the Treatment model (2.3) when fitted to treatment data with fixed δ, γ, and α.

Error level (%)	β(%)	ι(%)
0	2.01E−07	1.21E−09
1	0.07	0.13
5	0.40	0.65
10	0.76	1.20
20	1.35	2.38
30	2.17	4.04

Table 17
Monte Carlo simulation results: ARE for each parameter of the Treatment model (2.3) when fitted to treatment data with fixed α and δ.

Error level (%)	β(%)	γ(%)	ν(%)
0	0	1.78E−14	2.22E−14
1	0.11	1.70	0.14
5	0.54	7.99	0.71
10	1.02	18.30	1.33
20	1.80	32.61	2.97
30	3.88	85.33	5.51

Table 18
Monte Carlo simulation results: ARE for each parameter of the Treatment model (2.3) when fitted to cumulative incidence data with fixed α and δ.

Error level (%)	β(%)	γ(%)	ν(%)
0	0	1.78E−14	2.22E−14
1	4.79	68.85	13.53
5	7.04	124.46	31.22
10	5.85	102.67	37.83
20	7.58	95.46	49.05
30	12.40	105.26	53.70

Table 19
Monte Carlo simulation results: ARE for each parameter of the Treatment model (2.3) when fitted to cumulative incidence data with fixed α, γ and δ.

Error level (%)	β(%)	ν(%)
0	0	2.22E−14
1	0.30	1.16
5	1.59	6.19
10	2.87	10.92
20	5.35	20.81
30	8.43	34.36

Table 20
Monte Carlo simulation results: ARE for each parameter of the Treatment model (2.3) when fitted to treatment and cumulative incidence data.

Error level (%)	β(%)	α(%)	δ(%)	γ(%)	ν(%)
0	0	0	1.39E−14	1.78E−14	2.22E−14
1	5.05	97.88	26.55	19.56	8.54
5	22.88	104.43	147.09	20.92	39.88
10	34.41	135.74	4109.85	27.073	58.06
20	50.32	123.81	2947.64	24.69	105.67
30	54.94	143.09	5431.01	32.21	114.53

Table 21
Monte Carlo simulation results: ARE for each parameter of the Treatment model (2.3) when fitted to treatment and cumulative incidence data with fixed ν.

Error level (%)	β(%)	α(%)	δ(%)	γ(%)
0	0	0	1.39E−14	1.78E−14
1	0.87	90.55	23.98	18.10
5	4.60	95.84	26.43	19.46
10	7.71	83.55	26.06	17.64
20	19.97	97.50	62.654	22.98
30	26.99	113.38	3353.03	27.12

Table 22
Monte Carlo simulation results: ARE for each parameter of the Treatment model (2.3) when fitted to treatment and cumulative incidence data with fixed ν and δ.

Error level (%)	β(%)	α(%)	γ(%)
0	0	0	1.78E−14
1	0.86	8.86	2.07
5	4.23	44.21	9.98
10	6.47	64.21	14.14
20	12.26	109.53	22.42
30	15.49	127.93	28.56

Table 23
Summary of results. Structural and practical identifiability analysis of the models studied in the current paper.

Model	Modification	Prevalence/Treatment Data	Cumulative Data	Combined Data
SIR	Original	SI, PI	SI, PU	–
SEIR	Original	LSI, PU	SI, PU	–
	Fixed η	SI, PI	PU	–
	Fixed α	SI	PU	–
Treatment	Original	SU, PU	SU, PU	SLI, PU
	Fixed δ and γ	SI, PU	SI	–
	Fixed δ, γ and α	PI	–	–
	Fixed α and δ	PI	SI, PU	–
	Fixed ν	–	–	SI, PU

Notes. SI: Structurally Identifiable; LSI: Locally Structurally Identifiable; SU: Structurally Unidentifiable; PI: Practically Identifiable; PU: Practically Unidentifiable. The combined data for treatment model include both treatment and cumulative data.

and 4). On the other hand non of the models considered are practically identifiable from the cumulative incidence data. It is a concern of public health since, cumulative incidences are the main type of data reported by the health organizations. Even though prevalence data is rarely available, we continue our analysis with prevalence data. The SIR and SEIR model are structurally identifiable to prevalence data (see Proposition 1, and 2). We also examined the practical identifiability of the SIR and SEIR model by implementing the Monte Carlo simulations and sensitivity analysis. As a result, the SIR model is revealed to be practically identifiable from prevalence data, whereas the SEIR model is not. Furthermore, we analyzed the evolution of practical identifiability of SIR model from prevalence data, hence during an emerging infection projection of cases are estimated as the data becomes available. We found that SIR model is not practically identifiable from prevalence data at the initial stages of the outbreak. It becomes practically identifiable when the outbreak reaches its peak. This study shows that if possible, public health agencies should report prevalence instead of cumulative incidences, since it is the only data type which is practically identifiable. The parameters of the treatment model can not be identified from cumulative incidences or the number of treated individuals. It is straightforward from its structural unidentifiability that the treatment model is not practically identifiable to either cumulative incidence or treatment data. We investigate whether the treatment model would be structurally identifiable if both cumulative incidences and number of treated individuals are used (see Remark 3). The structural identifiability analysis reveals that the treatment model would be structurally identifiable if both data sets are used and if ν is fixed in the estimation procedure. But this is not the case in practice. When both treatment and cumulative data are used to estimate the parameters of the treatment model, it remains unidentifiable even when 2 out of 5 of its parameters are fixed (Table 22).

We found that a model could be structurally identifiable but not practically so. Specifically, given the cumulative incidence data, none of the aforementioned models are practically identifiable. In particular cases, modifying a model by further fixing specific parameters before estimating the remaining parameters could make the model practically identifiable by eliminating the possibility of indistinguishable parameters. Using Fisher Information Matrix, we computed the correlations between pairs of parameters in a model, given an observed state variable. If two parameters are highly correlated, such as β and η in the SEIR model with prevalence data, fixing one of them, say, η, reduces the dimension of the parameter space and makes the model identifiable (Table 9). However, this is not always the case, especially with cumulative incidence data. Because of the high inter-correlations among parameters in all models with cumulative incidence data, fixing one, two, and even three parameters still does not give identifiability to the remaining parameters (Table 19). Results of the analyses conducted in this study are summarized in Table 23.

The analyses performed in this paper shows that even though cumulative incidence data is the standard data type provided by the CDC and WHO, all outbreak models we analyzed are unidentifiable from this data structure. More generally, besides the complexity of the model itself, the type of data being used also has a significant effect on the parameter estimation problem. Therefore, we recommend carrying out an identifiability analysis before estimating multiple parameters in a complex model.

Acknowledgment

The authors N. Tuncer acknowledge support from the National Science Foundation (NSF) under grants DMS-1515442.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.mbs.2018.02.004.

References

[1] C.L. Althaus, Estimating the reproduction number of Ebola virus (EBOV) during the 2014 outbreak in West Africa, PLoS Curr. 6 (2014).
[2] H.T. Banks, S. Hu, W.C. Thompson, Modeling and Inverse Problems in the Presence of Uncertainty, CRC Press, 2014.
[3] G. Bellu, M.P. Saccomani, S. Audoly, L. D'Angio, DAISY: a new software tool to test global identifiability of biological and physiological systems, Comput. Methods Programs Biomed 88 (1) (2007) 52–61.
[4] A. Capaldi, S. Behrend, J. Smith, B. Berman, J. Wright, A.L. Lloyd, Parameter estimation and uncertainty quantification for an epidemic model, Math. Biosci. 9 (3) (2012) 553–576.
[5] J.D. Chapman, N.D. Evans, The structural identifiability of susceptible-infective-recovered type epidemic models with incomplete immunity and birth targeted vaccination, Biomed. Signal Process. Control 4 (4) (2009) 278–284.
[6] http://www.cdc.gov.
[7] G. Chowell, H. Nishiura, Transmission dynamics and control of Ebola virus disease (EVD): a review, BMC Med. 12 (196) (2014), http://dx.doi.org/10.1186/s12916-014-0196-0.
[8] O.T. Chris, J.R. Banga, E. Balsa-Canto, Structural identifiability of systems biology models: a critical comparison of methods, PLoS ONE 6 (11) (2011) 27755.
[9] M. Eisenberg, S. Robertson, J. Tien, Identifiability and estimation of multiple transmission pathways in cholera and waterborne disease, JTB 324 (2013) 84–102.
[10] N.D. Evans, L.J. White, M.J. Chapman, K.R. Godfrey, M.J. Chappell, The structural identifiability of the susceptible infected recovered model with seasonal forcing, Math. Biosci. 194 (2005) 175–197.
[11] D. Fisman, E. Khoo, A. Tuite, Early epidemic dynamics of the west african 2014 Ebola outbreak: estimates derived with a simple two-parameter model, PLoS Curr. Outbreaks, Sep. 18, Edition 1 (2014).
[12] B.R. Frieden, Science from Fisher Information, Cambridge University Press, New York, 2004.
[13] J.C. Lagarias, J.A. Reeds, M.H. Wright, P.E. Wright, Convergence properties of the Nelder–Mead simplex method in low dimensions, SIAM J. Optim. 9 (1) (1998) 112–147.
[14] L. Ljung, T. Glad, On the global identifiability of arbitrary model parametrizations, Automotica 30 (1994) 265–276.
[15] M. Lipsitch, et al., Transmission dynamics and control of severe acute respiratory syndrome, Science 300 (2003) 1966–1970, http://dx.doi.org/10.1126/science.1086616.
[16] M. Martcheva, An Introduction to Mathematical Epidemiology, Springer, New York, To appear.
[17] N. Meshkat, Z. Rosen, S. Sullivant, Algebraic tools for the analysis of state space models, arXiv:1609.07985.
[18] H. Miao, X. Xia, A.S. Perelson, H. Wu, On identifiability of nonlinear ODE models and applications in viral dynamics, SIAM Rev. 53 (1) (2011) 3–39.
[19] H. Pohjanpalo, System identifiability based on power-series expansion of solution, Math. Biosci. 41 (1978) 21–33.
[20] S. Riley, C.A. Donnelly, N. Ferguson, Robust parameter estimation techniques for stochastic within-host macroparasite models, J. Theor. Biol. 225 (2003) 419–430, http://dx.doi.org/10.1016/S0022-5193(03)00266-2.
[21] C. Fraser, Pandemic potential of a strain of influenza a (h1n1): early findings, Science 324 (2009) 1557–1561, http://dx.doi.org/10.1126/science.1176062.
[22] A.R. Tuite, J. Tien, M. Eisenberg, E. DJD, J. Ma, D. Fisman, Cholera epidemic in haiti, 2010 using a transmission model to explain spatial spread of disease and identify optimal control interventions, Ann. Intern. Med. 154 (2011) 593–601, http://dx.doi.org/10.7326/0003-4819-154-9-201105030-00334.
[23] N. Tuncer, H. Gulbudak, V. Cannataro, M. Martcheva, Structural and practical identifiability issues of immuno-epidemiological vector-host models, Bull. Math. Biol. 78 (9) (2016) 1796–1827.
[24] H.L.V. Trees, Detection, Estimation, and Modulation Theory, Part I, Wiley, New York, 1968.
[25] P.V.Den Driessche, W. James, Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission, Math. Biosci. 180 (1–2) (2002) 29–48, http://dx.doi.org/10.1016/s0025-5564(02)00108-6.
[26] http://www.who.int.