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.

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

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.

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

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.