Mathematical Biosciences and Engineering

https://www.aimspress.com/journal/mbe

MBE, 23(4): 913–939.
DOI: 10.3934/mbe.2026036
Received: 20 April 2025
Revised: 08 December 2025
Accepted: 14 January 2026
Published: 03 March 2026

Research article

Ambiguity in identifying parameters of an SIR model when fitting epidemic incidence data

B Shayak [1], Sana Jahedi [2] and James A Yorke [3,*]

[1] Department of Engineering, University of Maryland, College Park, MD, USA
[2] Department of Mathematics and Statistics, Oakland University, Rochester, MI, USA
[3] Department of Mathematics, University of Maryland, College Park, MD, USA

* **Correspondence:** Email: sb2344@cornell.edu.

Abstract: When a new pathogen emerges, determining the key transmission parameters plays a crucial role in formulating public health policies and controlling the spread of the pathogen. It is important to note that not every parameter is "identifiable". A parameter of a model is said to be globally structurally identifiable from some kind of perfect data if it can be determined uniquely. Whether that parameter is identifiable depends on what kind of perfect data is available. In this work, we developed a new mathematical concept, the "decay-growth ratio", and using it, we prove that the basic reproduction number is "globally structurally identifiable" from the knowledge of this ratio, and with a bit more information, the duration of infectiousness can be determined as well. That, however, assumes perfect, noise-free data which in reality is unattainable. A parameter is said to be "practically identifiable" if it can be reliably estimated from finite, noisy data. Practical identifiability is inherently dependent on both the nature of the available data and the inferential methodology employed. We proved that neither the basic reproduction number nor the duration of infectiousness is practically identifiable from the common summary statistics of an outbreak, specifically its mean, standard deviation, and amplitude. In fact, we showed that given any outbreak, and given any value greater than one for the basic reproduction, there is an SIR solution with the same mean and standard deviation as the outbreak's. We further demonstrated how this result can be extended to more complex multi-compartment epidemic models. Moreover, we provided indistinguishable fits to real epidemic data with extremely different parameter choices. This insensitivity of fit quality to parameter choices means that traditional curve-fitting cannot reliably infer the key outbreak parameters from case data alone. Taken together, our results highlight fundamental limits on estimating epidemic parameters from incidence data, i.e., the rate of new cases, or from prevalence data, i.e., the number of infected people at a given time.

Keywords: SIR model; basic reproduction number; decay-growth ratio; parameter estimation; identifiability; practical identifiability; structural identifiability

1. Introduction

When a new infectious disease emerges, determining key transmission parameters—such as the duration of infectiousness and the number of secondary infections caused by an infectious individual—can substantially influence the design of non-pharmaceutical interventions and the effectiveness of disease control efforts. Mathematical modeling has long been used to understand disease dynamics and predict key epidemiological factors [1–3]. Mathematical modeling of epidemics dates back to at least the 1910s when Sir Ronald Ross constructed a simple differential equation to model the spread of vector-borne disease malaria [4]. In 1927, William Kermack and Alexander McKendrick formulated a general mathematical theory of the spread of an epidemic [5]. Their formulation, now known as the SIR model, uses a system of nonlinear differential equations to describe how individuals move between the susceptible (S), infected (I), and recovered (R) compartments. They applied the model to an outbreak of plague in Bombay, British Empire (now Mumbai, India). In recent years, SIR models, sometimes with slight variations (e.g. Susceptible-Exposed-Infected-Recovered (SEIR), Susceptible-Infected-Susceptible (SIS) [6], Susceptible-Infected-Quarantined-Recovered-Susceptible (SIQRS) [7]), have been used to analyze a variety of epidemics, including outbreaks of measles, chickenpox, mumps and polio [1, 8, 9], Ebola virus [10–12], Zika virus [13–15], Nipah virus [16–18], and many applications to COVID-19 [19–21].

The values of R_0 and τ have been estimated for different strains of COVID-19 using different types of data such as case tracing of individuals by various consortia of scientists, including those from the London School of Hygiene & Tropical Medicine (LSHTM), Institute for Health Metrics and Evaluation (IHME), and Johns Hopkins University (JHU). Such measurements can be invaluable additions when fitting outbreaks with equations.

Although the mathematical modeling of epidemiology has a long and well-established history, parameter estimation remains a challenging task. During the COVID-19 pandemic, different research groups developed models and produced widely varying estimates for key transmission parameters, most notably the basic reproduction number. The basic reproduction number, R_0, is the expected number of secondary infections caused by a single infectious individual during their infectious period when almost everyone is susceptible [22, 23]. A systematic review by Dhungel et al. [24] reported that published R_0 estimates varied from 0.4 to 12.58.

In addition, when fitting a system of ordinary differential equations to a given data set, there are some implicit parameters that must be estimated, such as the initial conditions. Jahedi and Yorke [25] showed how four different research groups predicted different death tolls for New York City during early COVID-19, by considering different values for the initial susceptible fraction.

The problem of determining parameters in a system of equations from a given data set is called the identifiability problem [26]. Given a perfect, noise-free data set (such as a solution of a model), a parameter of a model is said to be globally structurally identifiable if it can be uniquely determined. A parameter of a model is said to be locally structurally identifiable if it can be estimated up to a finite number of distinct values given that perfect, noise-free data are available. When we say "structurally identifiable", we mean "global structural identifiability". Structural identifiability was first introduced by Bellman and Åström [27] in 1970. Since then, many different methods have been introduced in different fields to assess the structural identifiability of the parameters, and some of those methods have been applied to SEIR (or SIR) models [28–30]. Structural identifiability is an intrinsic property

of the model that also depends on the type of output data considered.

A model may be structurally identifiable with one type of output data, but becomes unidentifiable when another output data is considered [31]. In addition, for a given type of data, some parameters could be identifiable and others not. Evans et al. [28] proposed a transformation–based criterion for structural identifiability for a general form of an input-output system. They adapt their method to a form of SIR model with the prevalence data as the output and showed that the basic reproduction number and the duration of infectiousness is structurally identifiable, but the total population size is unidentifiable. Their method is complicated and requires one to find an infinitely differentiable transformation function that maps the trajectories of a system with a parameter vector p to those with a different vector p' while preserving the output.

In Section 3, we introduce a new simple theory that is purely dependent on the structure of the SIR model and show that the basic reproduction number and the duration of infectiousness are structurally identifiable. In practice, one could use the incidence curve (as the output data) to find unique values for the basic reproduction number and the duration of infectiousness. Our method is based on the ratio of the epidemic's final exponential decay rate to its initial exponential growth rate. We show that if the outbreak data is from a solution of our SIR model, then the parameters can be determined exactly. Specifically, let $S(t)$ and $I(t)$ be the fractions of the susceptible and infectious populations of an SIR solution. In practice, the available incidence or prevalence data represent only a portion of the actual cases. To represent an outbreak, we scale these fractions by a factor A so that AS and AI would be numbers of people. For example, AI might be the prevalence reported based on case reports. Assume the available incidence curve $\gamma(t)$ is equal to $-AS'(t)$ for some unknown scale factor $A > 0$ (where $A = \frac{\int \gamma(t)dt}{\int -S'(t)dt}$). Then the basic reproduction number, the duration of infectiousness, the infected fraction $\int -S'(t)dt$, and the scaling factor A are uniquely identifiable from the exponential rates alone, and these exponential rates are independent of A. Alternatively, if one instead has prevalence data and $AI(t)$ is known for some unknown A, again one can determine the same parameters because $I(t)$ has the same exponential rates as $-S'(t)$. Additionally, knowing the basic reproduction number allows one to determine the fraction of the total population infected during the outbreak.

In practice, accurately estimating the parameters of a model from the data requires more than just knowing that a model is structurally identifiable. The structural identifiability is applicable to perfect, noise-free data from a model, but in practice, such data is missing. The available data are often sparse and have some noise, and the model might not reflect the actual dynamics of the outbreak. For example, during the COVID-19 pandemic, the only data available were confirmed cases, and the total number of daily new cases was missing. Asymptomatic individuals played a major role in the dynamics of COVID-19, complicating the efforts of modelers to predict key transmission parameters [32] accurately. Practical identifiability is the type of identifiability that concerns whether the parameters of the model can be estimated from limited (finite) and noisy data in practice [33]. The parameters of a model may be structurally identifiable, but not practically identifiable. As we mentioned above, if the incidence data for the entire duration of the epidemic is known, using our decay-growth ratio, one could uniquely estimate the key transmission parameters. However, that data is missing. Similar to structural identifiability, many different methods have been developed to assess practical identifiability. For example, Sauer et al. [34], by applying the method of dynamical compensation [35], showed that if only incidence data prior to the epidemic peak were available, R_0 would not be practically identifiable.

In Section 4, we show that the three most basic statistics of an outbreak, the mean (μ), standard

deviation (σ), and total number of cases, are insufficient to determine the transmission parameters. We prove that for each value greater than one assigned to the basic reproduction number, there is an SIR solution whose outbreak has any desired choice of mean μ and standard deviation $\sigma > 0$. This illustrates that an unbounded range of parameter combinations can yield an outbreak with any desired statistics. We also extend our direct method to an SIR model with many compartments (see Theorem 3).

In Section 5, we numerically illustrate that the basic reproduction number and the duration of infectiousness are practically unidentifiable. For this section, we use two data sets, the daily incidence data of the Omicron (B1.1.529) variant of COVID-19 in New York City and London. A graphical representation of variant-specific outbreaks of COVID-19 for the two aforementioned cities is provided in Figure 1. Our focus on the Omicron outbreak is driven by its brevity and its high quality compared to other variants of COVID-19. Variant-specific testing and the S-gene target failure enabled the clear identification of B.1.1.529 cases. In Section 5, we analyze least squares fits of SIR solutions to these datasets. We find that despite widely differing values for the duration of infectiousness (e.g., 2.9 and 25.6 days in Figure 2, and 1.9 and 29.1 days in Figure 3), the fitted SIR curves are visually nearly identical. We illustrate this issue in Figure 2, which shows two nearly overlapping SIR fits to NYC's Omicron data, derived from very different parameter values. Similarly, Figure 3 illustrates two almost identical SIR fits to the London data, which are obtained from widely separated values for the duration of infectiousness. Because the SIR solutions are so similar, this least squares fit of the solutions to the data cannot be used to determine the parameters of the SIR model.

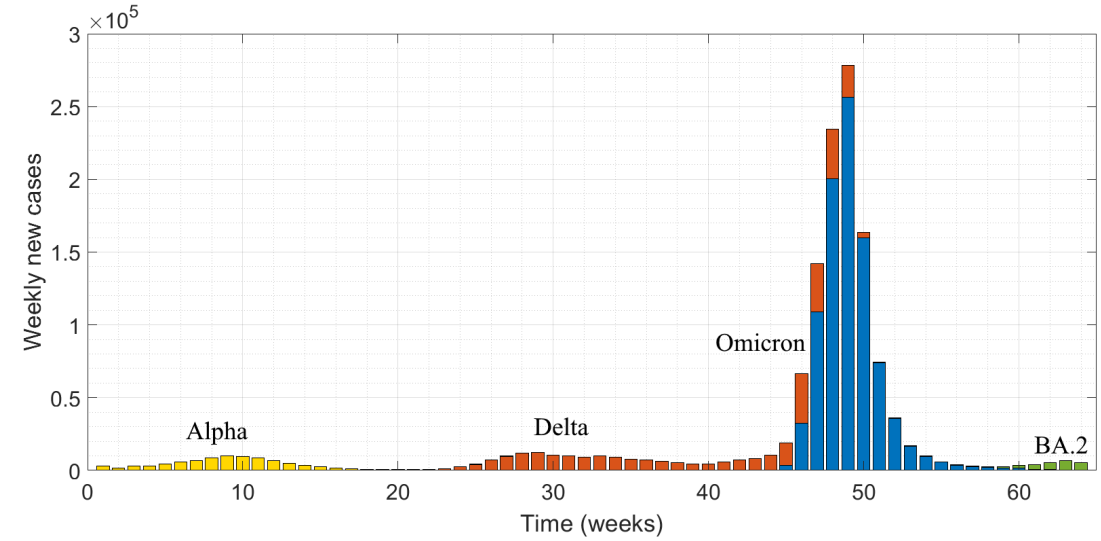
This paper investigates whether and how one can determine the transmission parameters, the basic reproduction number, and the duration of infectiousness from daily case count data, using the classical SIR model as the foundation. We aim to gain a deeper understanding of the properties and limitations of the SIR model.

Table 1. Estimated parameter values for SIR solution yield local best fit for NYC.

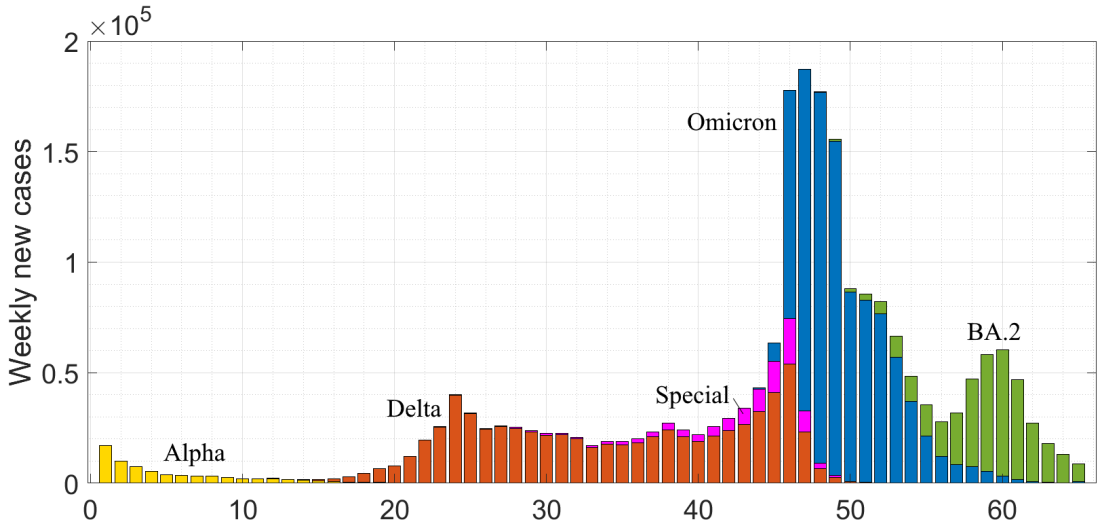
τ	ρ	A	I_0	Relative error	S_∞	λ_{growth}	λ_{decay}
2.9	1.523	1,709,000	$484/A$	0.064	0.414	0.1741	-0.1300
25.6	5.475	1,040,000	$1347/A$	0.059	0.004	0.1751	-0.0382

Table 2. Estimated parameter values for SIR solution yield local best fit for London data with holiday period excluded. To be precise, we define our objective function by optimizing the least square distance between the SIR model and the data only for day 1 through day 36 and from day 51 to day 84, excluding the data from day 37 to day 50. To make this clear, we have changed the color of data points that we exclude to red, see Section B in the Supplementary Material.

τ	ρ	A	I_0	Relative error	S_∞	λ_{growth}	λ_{decay}
1.9	1.280	2,421,000	0.00016	0.127	0.597	0.1474	-0.1241
29.1	5.344	1,009,000	0.0013	0.117	0.177	0.1493	-0.0019



(a)



(b)

Figure 1. Multi-strain COVID-19 weekly case totals for NYC and London. The data for New York City is shown in panel (a), and the data for London is shown in panel (b). The counts began when variant data became available, on January 24, 2021, for NYC and on January 31, 2021, for London. The last day is April 24, 2022, when the original B1.1.529 variant was almost completely extinguished. "Special" in panel (b) denotes the variant under investigation (VUI) 21-OCT-01, which spread at a rapid rate in the UK but was not seen in most other countries.

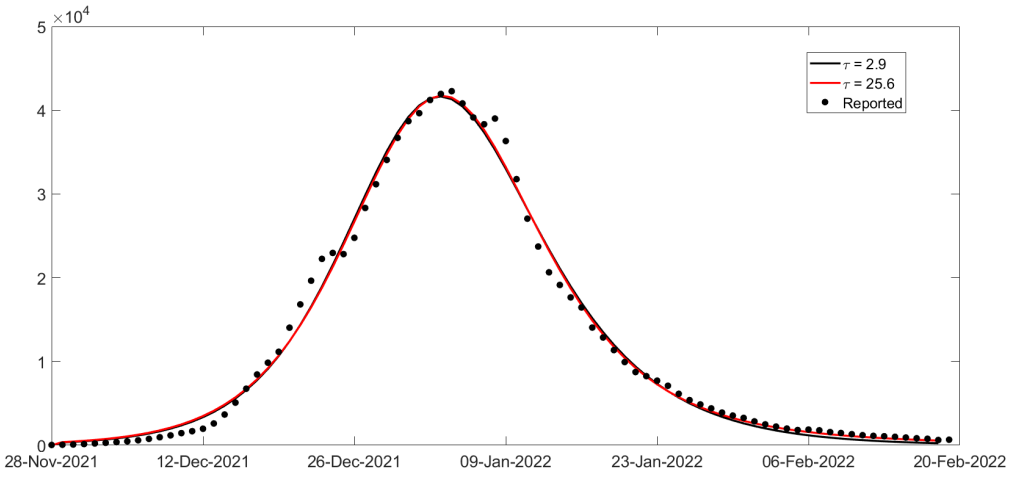


Figure 2. Fitting SIR solutions to NYC COVID-19 Omicron daily cases. The black dots show the smoothed daily new case reports of Omicron in New York City from November 28, 2021, to February 20, 2022. The solid black and red curves are the SIR solutions that are least squares fits to the data for different values of the average duration of infectiousness, $\tau = 2.9$ and 25.6. The corresponding ρ values are 1.505 and 5.482, respectively (see Section 5 and Table 1.) These two values of τ correspond to local best fits as τ is varied, with the second one (25.6) also being the global best fit. It may be hard to distinguish the two curves because the red curve lies almost on top of the black one—except that in the tail of the outbreak, they separate slightly. Because the SIR solutions are so similar, this least squares fitting of solutions to data cannot be used to determine the parameters of the SIR model, even within a reasonable range. As the graph illustrates, it is not possible to distinguish between $\tau = 2.9$ and $\tau = 25.6$.

2. Model and parameters

When the infected individuals are uniformly distributed in a population and the recovery and transmission rates are fixed, the outbreak (or epidemic) can be modeled using a standard SIR model :

$$\frac{dS}{dt} = -\beta S I, \tag{2.1a}$$

$$\frac{dI}{dt} = \beta S I - \frac{I}{\tau}, \tag{2.1b}$$

$$\frac{dR}{dt} = \frac{I}{\tau}. \tag{2.1c}$$

Model (2.1) above is described in many studies, including [36] and [37]. The variables S and I denote the fraction of the population that is susceptible and infected, respectively. The variable R denotes the fraction removed due to recovery or death. For each solution, $S + I + R$ is constant. In practice, this means that when S and I are known, R is also known. The average infectiousness per unit of time per person is β, and the average duration of infectiousness is denoted by $\tau > 0$. The basic reproduction number for Model (2.1) is $\beta\tau$. Herein, we write $\rho = \beta\tau$ and we refer to it as the basic reproduction

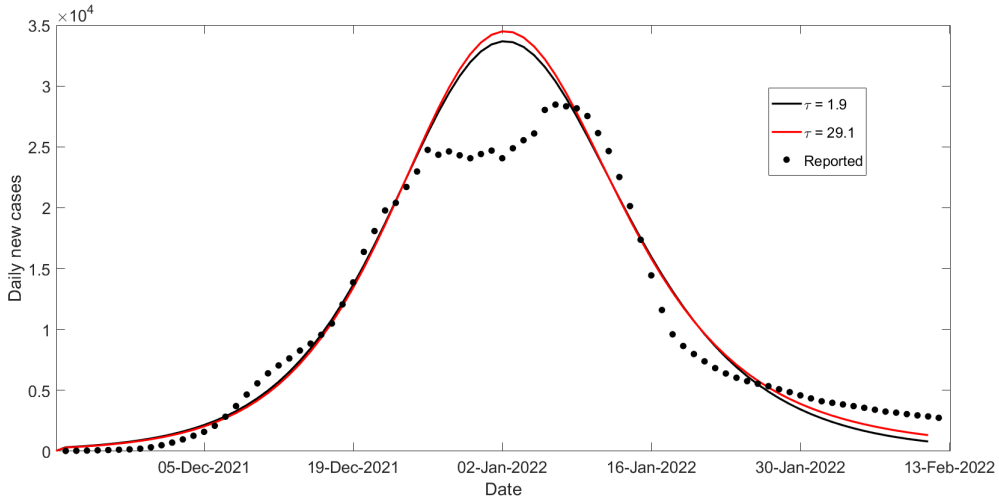


Figure 3. Fitting SIR solutions to London data with the holiday period excluded. Black dots denote centered 7-day average Omicron case counts in London. The black and red solid curves are two least squares fits of Model (2.2) to London data. As the plot shows, the two SIR solutions are quite similar and lie close to each other, even though their corresponding parameters (τ, ρ, I_{t_0}, A) are different, see Table 2 and its caption for details. The value of τ for each fit is chosen, and then the rest of the parameter values are determined using an optimization routine. The overall best least squares fit occurs for $\tau = 29.1$ for London Omicron data (*i.e.*, the smallest relative error) with a relative error of 11.7 percent and $\tau = 1.9$ is where we find a local minimum with a relative error of 12.7 percent. To see the relative error as a function of τ, see Figure A.2 in the supplementary material.

number. We recast Model (2.1) in terms of the parameter ρ, because ρ is the parameter that appears many times in our analysis below. Throughout this paper we assume $\rho > 1$. Our main model in this paper is the following

$$\frac{dS}{dt} = -\frac{\rho S I}{\tau}, \tag{2.2a}$$

$$\frac{dI}{dt} = \frac{\rho S I - I}{\tau}. \tag{2.2b}$$

Initializing our equations. For $\rho > 1$, the above system of equations has a solution:

$$I(t) + S(t) = \frac{1}{\rho}\ln(S(t)) + C,$$

where C is a constant. Since $I + S + R = C$, it follows the removed fraction is $R = -\frac{1}{\rho}\ln(S)$. We often denote time t by a subscript, writing $I(t)$ as I_t, $S(t)$ as S_t, $S(-\infty)$ as $S_{-\infty}$, and $S(\infty)$ as S_∞. Since $S(t)$ is monotonic and can be restricted to $(0, 1)$, we choose the maximum of S to be 1, *i.e.*,

$$S_{-\infty} = 1, \tag{2.3}$$

which implies $I_{-\infty} = 0$. Using these two limiting values at $-\infty$ to determine C yields $C = 1$. Hence, for all t,

$$I_t = -S_t + \frac{1}{\rho}\ln(S_t) + 1. \tag{2.4}$$

To initialize the system of equations, we use the above equation. Choose S_t to be some number slightly less than one (that is, $S_t = 1 - \varepsilon$ for some small $\varepsilon > 0$) and compute I_t. Since $\ln(1-\varepsilon) \approx -\varepsilon$ and $\frac{1}{\rho}\ln(1-\varepsilon) \approx -\frac{1}{\rho}\varepsilon$, we set

$$I_t \approx -(1-\varepsilon) - \frac{1}{\rho}\varepsilon + 1 = \varepsilon(1 - \frac{1}{\rho}) \text{ for } \varepsilon << 1. \tag{2.5}$$

In this case, the fraction of "removed" is $R_t = 1 - S_t - I_t \approx \frac{\varepsilon}{\rho}$. More precisely, $R_t = -\frac{1}{\rho}\ln(S_t)$.

Remark. Alternatively, many authors prefer to assume that at some time t_0, $S_{t_0} = 1 - I_{t_0} = 1 - \varepsilon$ (where ε is small). That is, $I_{t_0} = \varepsilon$, in which case $C = 1 + \varepsilon/\rho$. For an outbreak in a large city of 10 million people with one infected person, $\varepsilon = 10^{-7}$. Suppose, for example, $\rho = 2$. Then $C \approx 1 + \ln(S_{t_0})/\rho \approx 1 - 0.00000005$.

Instead, in this paper, we choose $C = 1$. If we were to choose a starting point, it might be day 4 of our data. We smoothed our NYC and London data to remove strong weekly periodic fluctuations. We use a centered seven-day moving average. That means that day 4, with 135 reported cases, is actually the first point that was the center of seven consecutive points and the first to receive our smoothing.

3. The decay-growth ratio $Rat(S)$ for identifying transmission parameters

Let $\gamma(t)$ be the rate of new cases for an outbreak of an infection and assume $\gamma(t) \to 0$ as $t \to \pm\infty$. At the beginning of an outbreak, new cases are likely to increase exponentially, while at the end, new cases decrease exponentially. This section considers situations in which both asymptotic rates exist. Here we examine the ratio of those two rates.

The outbreak's initial exponential growth rate is defined as

$$\lambda_{\text{growth}}(\gamma) := \lim_{t\to-\infty} (d/dt)\log(\gamma(t)). \tag{3.1}$$

The outbreak's final exponential decay rate is defined as

$$\lambda_{\text{decay}}(\gamma) := \lim_{t\to+\infty} (d/dt)\log(\gamma(t)). \tag{3.2}$$

We may expect $\lambda_{\text{decay}} < 0$ and $\lambda_{\text{growth}} > 0$, and for $t << 0$, $\gamma(t) \approx c_1 e^{\lambda_{\text{growth}} t}$, and for $t >> 0$, $\gamma(t) \approx c_2 e^{\lambda_{\text{decay}} t}$ for some positive constants c_1, c_2. For any outbreak for which these exponents exist, we define

$$Rat(\gamma) := |\lambda_{\text{decay}}|/\lambda_{\text{growth}}. \tag{3.3}$$

When $S = S(\cdot)$ is an SIR solution, we will abuse the notation and write $Rat(S(\cdot))$ for $Rat(-S')$.

The following theorem shows that ρ and τ are globally structurally identifiable.

Theorem 1 (Identifying ρ and τ)**.** *Let $S = S(\cdot)$ be an SIR solution of Model* (2.2) *satisfying Eq.* (2.3) *with transmission parameters ρ and τ. Given $\lambda_{growth}(S(\cdot))$ and $\lambda_{decay}(S(\cdot))$, τ and ρ can be uniquely identified.*

We will use Propositions 1 and 2 in the proof of Theorem 1.

The following proposition gives closed formulas for λ_{growth}, λ_{decay}, and Rat for the incidence curve of an SIR solution. The formulas show that the rates λ_{growth} and λ_{decay} can be derived from knowing

the model, and as part (P_2) of the following proposition shows, λ_{growth} and λ_{decay} do not depend on a specific data. They are intrinsic properties of the model; regardless of the type of data used, the initial exponential growth and the final exponential decay rates are the same and depend solely on the model, not on a specific initial condition or data.

We show in this proposition that $-S'$ and I have the same growth and decay exponential rates, so both yield the same Rat.

Proposition 1. *Let $S(\cdot), I(\cdot)$ be the SIR solution in the above theorem. Then:*

(P_1)

$$\lambda_{growth}(S(\cdot)) = \frac{\rho - 1}{\tau} > 0, \tag{3.4}$$

$$\lambda_{decay}(S(\cdot)) = \frac{\rho S_\infty - 1}{\tau} < 0, \tag{3.5}$$

$$\text{Rat}(S(\cdot)) = \frac{1 - \rho S_\infty}{\rho - 1}. \tag{3.6}$$

(P_2) $S(\cdot)$ and $I(\cdot)$ have the same growth and decay rates and so the same Rat.

Proof.

$$\lambda_{\text{decay}} = \lim_{t\to+\infty} (\text{d/dt}) \log(-S') = (\log S')'(+\infty).$$
$$\lambda_{\text{growth}} = \lim_{t\to-\infty} (\text{d/dt}) \log(-S') = (\log S')'(-\infty).$$

$$(\log S')' = \frac{S''}{S'} = \frac{\left(\frac{\rho}{\tau}\right)[S'I + SI']}{\rho SI/\tau} = \frac{S'}{S} + \frac{I'}{I}.$$

$\frac{S'}{S} = -\frac{\rho I}{\tau} \to 0$ as $t \to \pm\infty$, and $\frac{I'}{I} = \frac{\rho S - 1}{\tau} \to \frac{\rho S_{\pm\infty} - 1}{\tau}$ as $t \to \pm\infty$, which equals $\frac{\rho S_\infty - 1}{\tau}$ as $t \to \infty$ and equals $\frac{\rho - 1}{\tau}$as $t \to -\infty$ since $S_{-\infty} = 1$. Therefore Eqs. (3.4) and (3.5) are satisfied. By definition, $Rat(S(\cdot)) = |\lambda_{\text{decay}}|/\lambda_{\text{growth}}$, which yields Eq. (3.6).

It is obvious that when $\rho > 1$, λ_{growth} is positive. Write S_{peak} for the value of S at which $dI/dt = 0$ and $I \neq 0$. Therefore $S_{\text{peak}} = 1/\rho$ and $S_{peak} > S_\infty$, so $\rho S_\infty - 1 < 0$ where $0 < S_\infty < 1$. Hence, $\lambda_{\text{decay}} < 0$.

Note that $(\log S')'(\pm\infty) = \frac{I'}{I}(\pm\infty)$, proving (P_2). □

Proposition 2. *Let S be as assumed in Theorem 1. There exists a one-to-one correspondence between the values of ρ and the values of Rat.*

To prove Proposition 2, we need the following two lemmas, in which we prove two facts: 1- The basic reproduction number ρ is a decreasing function of the final size S_∞, which is proven in Lemma 1. 2- *Rat* is an increasing function of S_∞, this is proven in Lemma 2.

Setting $t = \infty, S_t = S_\infty$, and $I_\infty = 0$ in Eq. (2.4) provides a well-known relationship between ρ and S_∞ [36, 38]:

$$\rho = \frac{\ln(S_\infty)}{S_\infty - 1}. \tag{3.7}$$

The following lemma shows that the above equation represents a monotonic relationship between ρ and S_∞.

Lemma 1. *Define* $H(x) := \dfrac{\ln(x)}{x-1}$. $H : (0,1) \to (1,+\infty)$ *is a strictly decreasing and onto function. Furthermore,* $H(x) \to 1$ *as* $x \to 1$.

Proof. Differentiating $H(x)$ with respect to x yields

$$H'(x) = \frac{\frac{1}{x}(x-1) - \ln(x)}{(x-1)^2} = \frac{1 - \frac{1}{x} - \ln(x)}{(x-1)^2}. \tag{3.8}$$

To analyze the sign of $H'(x)$, define the auxiliary functions

$$f_1(x) = 1 - \frac{1}{x} \quad \text{and} \quad f_2(x) = \ln(x),$$

both defined on the interval $(0,1)$. $f_1'(x) - f_2'(x) = \dfrac{1-x}{x^2}$. Hence, $f_1'(x) - f_2'(x)$ is strictly positive for all $x \in (0,1)$. Therefore, $f_1(x) - f_2(x)$ is strictly increasing on $(0,1)$. Since $f_1 - f_2$ is strictly increasing and equals zero at $x = 1$, it follows that

$$f_1(x) - f_2(x) < 0 \quad \text{for all} \quad x \in (0,1).$$

Consequently, $H'(x) < 0$ for all $x \in (0,1)$. Therefore, H strictly decreases in $(0,1)$. Finally, note that

$$\lim_{x\to 0^+} H(x) = +\infty \quad \text{and} \quad \lim_{x\to 1^-} H(x) = 1.$$

Since H is a continuous function and is strictly decreasing on $(0,1)$, the intermediate value theorem implies that H maps onto $(1,+\infty)$, as required. □

The following lemma describes that $\mathrm{Rat}(S(\cdot)) = \frac{1-\rho S_\infty}{\rho - 1}$ is a monotonic function of S_∞. Note that by Eq (3.7), $\rho = \dfrac{\ln(S_\infty)}{S_\infty - 1}$.

Lemma 2. *Define* $G(x) = \dfrac{1 - H(x)\,x}{H(x) - 1}$ *where* $H(x) = \dfrac{\ln(x)}{x-1}$. *Function* $G(x)$ *is a strictly increasing and onto function.*

Proof. From Lemma 1, $H(x) \neq 1$ for $x \in (0,1)$. Differentiating $G(x)$ yields

$$G'(x) = \frac{H'(x)\,(x-1) + H(x) - H(x)^2}{(H(x)-1)^2}. \tag{3.9}$$

By Eq. (3.8), $(x-1)H'(x) = \dfrac{1}{x} - H(x)$. Hence,

$$G'(x) = \frac{\frac{1}{x} - H(x) + H(x) - H(x)^2}{(H(x)-1)^2} = \frac{\frac{1}{x} - H(x)^2}{(H(x)-1)^2}.$$

Substituting the expression for $H(x) = \dfrac{\ln(x)}{x-1}$, we rewrite the derivative as

$$G'(x) = \frac{(x-1)^2 - x\,(\ln(x))^2}{(H(x)-1)^2\,x\,(x-1)^2}.$$

Note that, again, the denominator is non-zero for $x \in (0, 1)$. Define the auxiliary function

$$Q(x) = (x-1)^2 - x\,(\ln(x))^2,$$

so that

$$G'(x) = \frac{Q(x)}{(H(x)-1)^2\, x\,(x-1)^2}.$$

The sign of $G'(x)$ depends entirely on the sign of $Q(x)$. We now show that $Q(x) > 0$ for all $x \in (0, 1)$. Differentiating $Q(x)$ results in

$$Q'(x) = 2(x-1) - 2\ln(x) - (\ln(x))^2,$$

$$Q''(x) = 2\left(1 - \frac{1}{x} - \frac{\ln(x)}{x}\right),$$

$$Q'''(x) = 2\frac{\ln(x)}{x^2}.$$

Observe that $\ln(x) < 0$ for all $x \in (0, 1)$, so $Q'''(x) < 0$ on $(0, 1)$. Therefore, $Q''(x)$ is strictly decreasing on $(0, 1)$. Since $Q''(x)$ is strictly decreasing and $Q''(1) = 0$, it follows that $Q''(x) > 0$ for all $x \in (0, 1)$. Consequently, $Q'(x)$ is strictly increasing on $(0, 1)$. Since $Q'(x)$ is strictly increasing and $Q'(1) = 0$, it follows that $Q'(x) < 0$ for all $x \in (0, 1)$. Thus, $Q(x)$ is strictly decreasing on $(0, 1)$. Since $Q(x)$ is strictly decreasing and $Q(1) = 0$, it follows that $Q(x) > 0$ for all $x \in (0, 1)$. Therefore, $G'(x) > 0$ for all $x \in (0, 1)$, and thus G is strictly increasing on $(0, 1)$. Finally, note that

$$\lim_{x\to 0^+} G(x) = 0 \quad \text{and} \quad \lim_{x\to 1^-} G(x) = 1.$$

Since G is continuous and strictly increasing on $(0, 1)$, the intermediate value theorem implies that G maps $(0, 1)$ onto $(0, 1)$, as required. □

Proof of Proposition 2. To prove the proposition, we show that there is a one-to-one, strictly increasing, continuous function $F : [0, 1] \to [0, 1]$ satisfying $F(0) = 0$ and $F(1) = 1$ such that $\frac{1}{\rho} = F(Rat)$ for $Rat \in (0, 1)$. We establish the proof by showing that F is the composition of two strictly monotonic "onto" functions from $(0, 1)$ onto $(0, 1)$, and therefore F is strictly monotonic.

Lemma 2 shows that $Rat = G(S_\infty)$ is a strictly increasing function of S_∞ that maps (0,1) onto (0,1). Hence, $G^{-1}(Rat)$ is strictly increasing and maps (0,1) onto (0,1).

Lemma 1 shows that $\rho = H(S_\infty)$ is a strictly decreasing function of S_∞ that maps (0,1) onto $(1, \infty)$. Hence, $1/H$ is strictly increasing and maps $(0, 1)$ onto $(0, 1)$. Let

$$F = \frac{1}{\rho} = 1/H(G^{-1}(Rat)). \tag{3.10}$$

The composition of two strictly increasing functions is strictly increasing, hence, F is a strictly increasing function that maps $(0, 1)$ onto $(0, 1)$. □

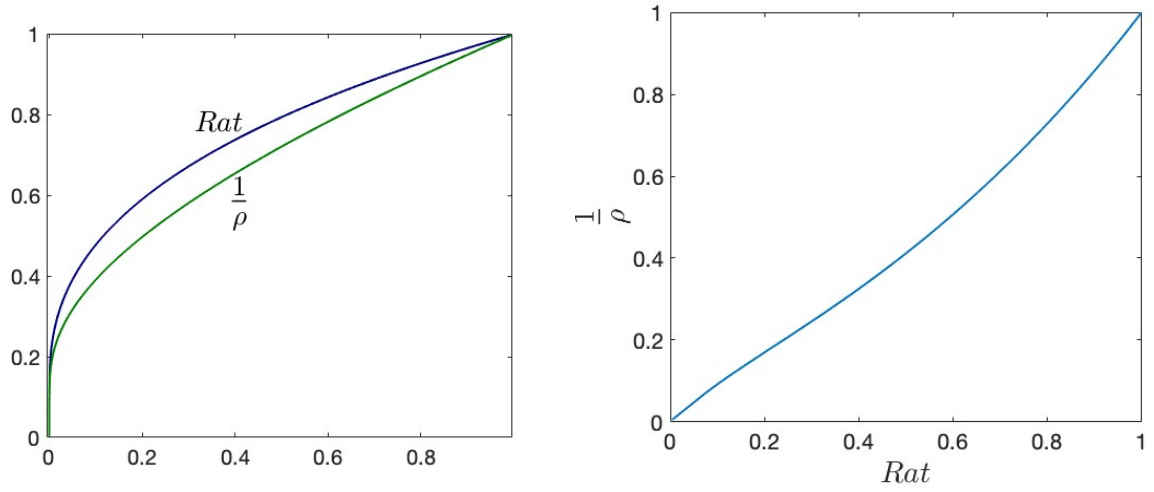
Proof of Theorem 1. Given λ_{growth} and λ_{decay}, $Rat(S(.)) = |\lambda_{\text{decay}}|/\lambda_{\text{growth}}$ can be uniquely calculated. By Proposition 2, ρ can be uniquely calculated from $\rho = 1/F = H(G^{-1}(Rat))$.

Once ρ is identified, τ can be calculated from either of the following equations.

$$\tau = \frac{\rho - 1}{\lambda_{\text{growth}}} \text{ or } \tau = \frac{\rho S_\infty - 1}{\lambda_{\text{decay}}} \tag{3.11}$$

Note that both of the above equations will yield the same value for τ, because by Eq. (3.3), $Rat(S(.)) = |\lambda_{\text{decay}}|/\lambda_{\text{growth}}$ and by Eq. (3.6), $Rat(S(.)) = \frac{1-\rho S_\infty}{\rho-1}$, which means $\frac{-\lambda_{\text{decay}}}{\lambda_{\text{growth}}} = \frac{1-\rho S_\infty}{\rho-1}$. That is equivalent to $\frac{\rho-1}{\lambda_{\text{growth}}} = \frac{1-\rho S_\infty}{-\lambda_{\text{decay}}}$. □

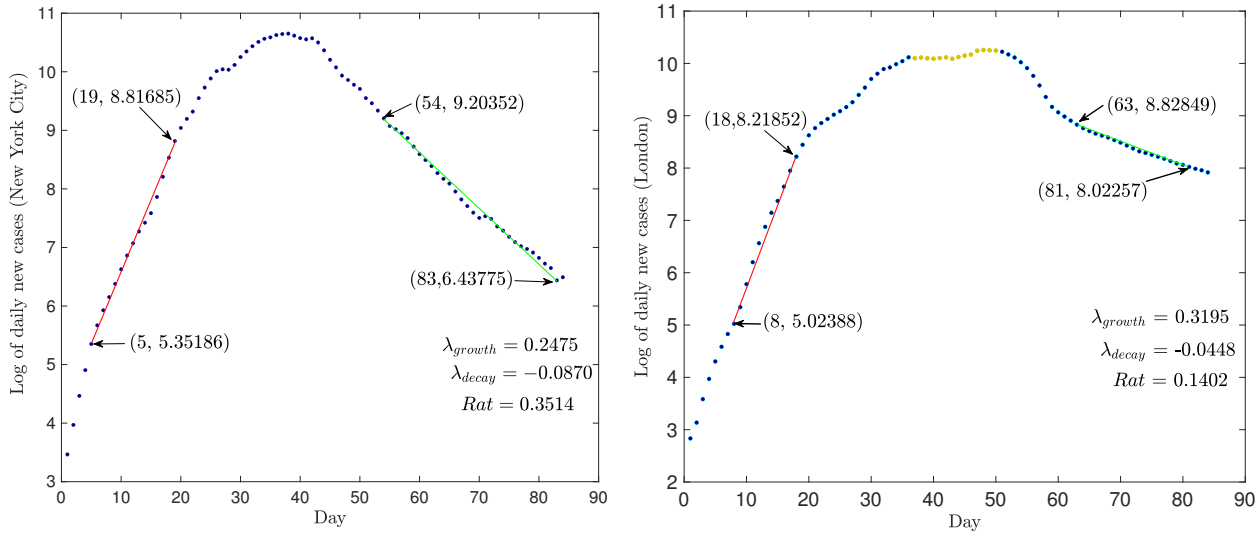
Plotting the basic reproduction number as a function of *Rat*. The graph of F (Eq. (3.10)) is the set of pairs of the form $(Rat, F(Rat)) = (Rat, \frac{1}{H(G^{-1}(Rat))})$, and that is the same as the parametric graph $(Rat(S_\infty), \frac{1}{\rho(S_\infty)})$ for $S_\infty \in (0, 1)$. The parametric graph is easy to plot, see Figure 4.



(a) *rat* and $1/\rho$ as functions of $x = S_\infty$ **(b)**

Figure 4. *Rat* uniquely identifies ρ. For the SIR model, each final susceptible value $S_\infty \in (0, 1)$ corresponds to unique values of *Rat* and ρ, which is demonstrated in panel (a). One could plot ρ as a function of *Rat*, by parameterization in the terms of S_∞, i.e., plotting pairs $(Rat(S_\infty), \dfrac{1}{\rho(S_\infty)})$ for each $S_\infty \in (0, 1)$, this is illustrated in panel (b).

To identify ρ and τ for New York City and London using the early exponential growth and the late exponential decay, we first plot the daily new cases data on a log scale, as shown in Figure 4. The slopes of the curves in the log-scale plots are the exponents we use (even though they are not limits taken at $\pm\infty$). Given the values of λ_{growth} and λ_{decay}, we can calculate *Rat* using Eq. 3.6. Once *Rat* is known, we can numerically solve the nonlinear equation $Rat = \frac{(S_\infty-1)-S_\infty \ln(S_\infty)}{\ln(S_\infty)-(S_\infty-1)}$ for S_∞. Once S_∞ is known, we can solve Eq. (3.7) for ρ. Finally, we can plug in either of the equations given in Eq. (3.11) to solve for τ. Note that, as was mentioned in the proof of Theorem 1, both yield the same value for τ if infinite precision is used. The values estimated from the incidence data given in Figure 5 are presented in Table 3.



(a) **(b)**

Figure 5. Daily cases from New York City and London in log scale. When data are plotted in log scale for these two cities, the early and late parts of the data turn out to be nearly straight lines. That corresponds to exponential growth and exponential decay.

Table 3. Key transmission parameters estimated from incidence data in Figure 5, using *Rat* method.

City	λ_{decay}	λ_{growth}	*Rat*	S_∞	ρ	τ
NYC	-0.0870	0.2475	0.3514	0.03366	3.5096	10.1397
London	-0.0448	0.3195	0.1402	0.0002997	8.1153	22.2701

4. Fitting SIR solutions to a mean, standard deviation, and amplitude

In the previous section, we showed that the parameters ρ and τ are globally structurally identifiable when the output data are perfect, noise-free data viz. the incidence data $(-S'(t))$; see Theorem 1. However, the identifiability of the parameters depends critically on the type of output data that is provided. In this section, we prove that when the type of output data is limited and only the mean, standard deviation, and amplitude of an outbreak are provided, ρ and τ cannot be uniquely estimated. In fact, we will show that for each $\rho > 1$, there exists an SIR solution with that mean and standard deviation, and it can be adjusted to give the desired amplitude.

Below, we define several new concepts that will be used throughout this section.

Definition 1 (Target triple)**.** Let $f : \mathbb{R} \to [0, \infty)$ be an integrable function. Assume

$$A_f := \int_{\mathbb{R}} f(t)dt \in (0, \infty), \tag{4.1}$$

$$\mu_f := \frac{\int_{\mathbb{R}} tf(t)dt}{\int_{\mathbb{R}} f(t)dt} = \frac{\int_{\mathbb{R}} tf(t)dt}{A_f} \in \mathbb{R}, \text{ and} \tag{4.2}$$

$$\sigma_f := \Big(\frac{\int_{\mathbb{R}}(t-\mu_f)^2 f(t)dt}{A_f}\Big)^{1/2} \in (0,\infty). \tag{4.3}$$

In other words, let A_f, μ_f, and σ_f denote the size of f, the mean time, and the standard deviation of a given function f (where usually $f = -S'$). We call such a triple (A_f, μ_f, σ_f) a **target triple**.

Definition 2 (3-statistic fit)**.** For a target triple (A_f, μ_f, σ_f), we say a function $g : \mathbb{R} \to [0,\infty)$ is a **3-statistic fit** (for the target triple) if $(A_g, \mu_g, \sigma_g) = (A_f, \mu_f, \sigma_f)$.

Sometimes we write Model (2.2; ρ, τ) to indicate which values are assigned to parameters of Model (2.2). Theorem 2 says that for each $\rho \in (1,\infty)$, there is a 3-statistic fit. A bit of caution is appropriate: a "3-statistic fit" does not necessarily mean a high-quality fit. The standard for the best fit is the least squares fit, which may differ significantly from a 3-statistic fit; see Section 5 for least squares fits.

Theorem 2. Let $\rho > 1$. Let (A_f, μ_f, σ_f) be a target triple. There exist $A > 0$, $\tau > 0$, and a non-constant solution (S_1, I_1) of Model (2.2; ρ, τ) such that the resulting $-S_1'$ multiplied by A is a 3-statistic fit to (A_f, μ_f, σ_f).

First, we establish some useful facts.

Proposition 3. Assume $\rho > 1$ and that $(S(t), I(t))$ is a non-constant solution of Model (2.2; $\rho, \tau = 1$). The following are true:

(I) $-S'$ has a finite mean time μ, which, without loss of generality, we may assume to be zero.
(II) For every $a > 0$ and μ, $(S(\frac{t-\mu}{a}), I(\frac{t-\mu}{a}))$ is a solution of Model (2.2; $\rho, \tau = a$).
(III) Let $f(t) = -S'(\frac{t-\mu}{a})$. Its mean time is μ, and its standard deviation is $\sigma \cdot a$, where σ denotes the standard deviation of $-S'(t)$.

Proof. To prove (I), observe that initially near $(S, I) = (1, 0)$, we have

$$\frac{dI}{dt} \approx \frac{\rho - 1}{\tau} I.$$

So $I(t)$ initially grows exponentially. Eventually $I(t)$ is decreasing, so $\frac{(\rho S(t)-1)}{\tau} < 0$ and ρS is decreasing. Therefore, $I(t)$ decreases exponentially as $t \to \infty$. Since $S' = -\rho S I/\tau$ and S is bounded, it follows that $|S'(t)| \to 0$ exponentially as $t \to \pm\infty$. It follows that the mean $S'(t)$ is finite.

To prove statement (II), note that since $(S(t), I(t))$ is a solution of Model (2.2; $\rho, \tau = 1$), then by Eq. (2.2a),

$$S'(t) = \frac{d}{dt} S(t) = -\rho S(t) I(t).$$

Therefore $(S(\frac{t-\mu}{a}), I(\frac{t-\mu}{a}))$ satisfies

$$\frac{d}{dt} S\Big(\frac{t-\mu}{a}\Big) = \frac{S'(\frac{t-\mu}{a})}{a} = \frac{-\rho S(\frac{t-\mu}{a}) I(\frac{t-\mu}{a})}{a},$$

so $S(\frac{t-\mu}{a})$ satisfies Eq. (2.2a) with $\tau = a$. Similarly, $I(\frac{t-\mu}{a})$ satisfies Eq. (2.2b) when $\tau = a > 0$. Furthermore, Eq. (2.4) is satisfied, proving that $(S(\frac{t-\mu}{a}), I(\frac{t-\mu}{a}))$ is a solution of Model (2.2; $\rho, \tau = a$).

To prove statement (III) about the mean, we apply a change of variables $x = \frac{t-\mu}{\tau}$ or $t = x\tau + \mu$.

$$\mu_f = \frac{\int_{\mathbb{R}} tf(t)dt}{\int_{\mathbb{R}} f(t)dt} = \frac{\int_{\mathbb{R}} tS'(\frac{t-\mu}{\tau})dt}{\int_{\mathbb{R}} S'(\frac{t-\mu}{\tau})dt} \tag{4.4}$$

$$= \frac{\int_{\mathbb{R}} (x\tau+\mu)S'(x)\tau dx}{\int_{\mathbb{R}} S'(x)\tau dx} = \tau\frac{\int_{\mathbb{R}} xS'(x)dx}{\int_{\mathbb{R}} S'(x)dx} + \mu = \tau\cdot 0 + \mu = 0 + \mu. \tag{4.5}$$

The standard deviation assertion in (III) follows from a similar calculation and is omitted here. □

Proof of Theorem 2. By the assumption of the theorem, ρ is greater than one. Therefore, there exists a non-constant solution $(S(t), I(t))$ of Model (2.2; $\rho, \tau = 1$). By part (I) of Proposition 3, $-S'(t)$ has a finite mean time. Then we can choose $(S(t), I(t))$ so that $-S'(t)$ has a mean time of 0. Let (A_f, μ_f, σ_f) be a target triple. By Proposition 3, $(S(\frac{t-\mu_f}{\tau}), I(\frac{t-\mu_f}{\tau}))$ is a solution of Model (2.2; ρ, τ) such that $-S'(\frac{t-\mu_f}{\tau})$ has a mean time of μ_f and a standard deviation of $\sigma_1 \cdot \tau$, where σ_1 is the standard deviation of $-S'(t)$.

Let $\tau = \frac{\sigma_f}{\sigma_1}$. Let $(S_1(t), I_1(t)) = (S(\frac{t-\mu_f}{\tau}), I(\frac{t-\mu_f}{\tau}))$. Then $(S_1(t), I_1(t))$ is a solution of Model (2.2; ρ, τ), and its incidence has a mean time of μ_f and a standard deviation of σ_f.

Let $A = \frac{A_f}{-\int_{\mathbb{R}} S_1'(t)dt}$. Then $-AS_1'(t)$ has integral equal to A_f and $-AS_1'$ is a 3-statistic fit for the target triple. □

4.1. Even more 3-statistic fits when using multicompartmental SIR models

Is it possible that Model (2.2) has so many 3-statistic fits because it is too simple? We show in the following that many more complex models have far more 3-statistic fits to a given target triple.

More complex models allow splitting the population into subgroups, each having its own susceptible and infected fractions and infectious period.

For example, suppose the population is divided into five age categories, with each category having two levels of susceptibility. In addition, each of these categories is divided into three subcategories based on the number of social contacts they have. That means that in this somewhat arbitrary example, there are $N = 5 \times 2 \times 3$ compartments.

In general, let N be the number of subpopulations (or compartments). The susceptible and infectious populations are represented by N-vectors whose components are fractions of the total population: $\vec{S} = (S_1, \ldots, S_N)$ and $\vec{I} = (I_1, \ldots, I_N)$. We will let i, j be coordinates, $1 \le i, j \le N$. Each S_i (or I_i) is the fraction of group i that is susceptible (infectious, respectively) and $S_i(-\infty) = 1$. The mean duration of infectiousness of I_i is denoted by τ_i. There are N^2 basic reproduction numbers $\rho_{i,j} > 0$, the number of people in group i who will be potentially infected by an infected person in group j, and $S_i\frac{\rho_{i,j}}{\tau_j}I_j$ is the fractional rate of new cases in group i caused by infectious people in group j. Typically such data are not collected and available for different categories; if it were collected for N categories, the model would have $N \times (N+1)$ unknown parameters $\rho_{i,j}$ and τ_i.

We do not explicitly specify the relative sizes of the groups. Each number $\rho_{i,j}$ depends in part on the sizes of the groups i and j, and because they are transmission rates, they are extremely difficult to estimate. Errors in these numbers might be systematically high (or low), and the errors would compound and not cancel each other. The SIR model, which describes the transmission of infection, is

as follows.

$$\frac{d}{dt}S_i = -S_i \sum_{j=1}^{N} \frac{\rho_{i,j}}{\tau_j} I_j, \tag{4.6a}$$

$$\frac{d}{dt}I_i = S_i \sum_{j=1}^{N} \frac{\rho_{i,j}}{\tau_j} I_j - \frac{I_i}{\tau_i}, \text{ where} \tag{4.6b}$$

$$S_i(-\infty) = 1, \quad i = 1, \ldots, N. \tag{4.6c}$$

Theorem 3 below is an analog of Theorem 2, with many free transmission parameters, analogs of ρ and τ.

Although $\rho > 1$ is needed in Theorem 2, for Model (2.2) to have an outbreak, we do not have lower bounds for the values of N^2 $\rho_{i,j}$, since an outbreak can also depend on the relative sizes of τ_i. It is sufficient to have $\rho_{i,i} > 1$ for some i to have an outbreak, since then the i^{th} group would support an outbreak by itself.

Let $p_i > 0$ be the fraction of the total population in group i, so $\sum p_i = 1$. When using a solution $(\vec{S}, \vec{I})$ to fit an outbreak, we use the fractional rate of new cases, namely $-\sum_i p_i S_i'$. For $A > 0$, we say $g(t) := -A \sum_i p_i S_i'$ is a **3-statistic fit** to (A_f, μ_f, σ_f) if $\int_{-\infty}^{\infty} g = A_f$, the mean and the standard deviation of g are μ_f and σ_f.

For cases where $N = 30$ as described above, there is a 900-dimensional set of matrices M that yield 3-statistic fits to each target. By the Perron-Frobenius theorem [39], the eigenvalue with the largest absolute value for M is a real positive eigenvalue. Therefore, if M does not have a real eigenvalue > 1, there is no outbreak, so there is no 3-statistic fit, since then each solution of Model (4.6) has $I(t) = 0$ for all t.

For simplicity, to specify which values of $(\rho_{i,j})$ and τ_i a solution of Model (4.6) depends on, we write $(\vec{S}, \vec{I})(t)$ as $(\vec{S}, \vec{I})(t; (\rho_{i,j}); \tau_1, \ldots, \tau_N)$, or that it is a solution of Model (4.6; $(\rho_{i,j})$; $\tau_1, \ldots, \tau_N$).

Theorem 3 (An N^2-dimensional set of 3-statistic fits to each target triple)**.** Let (A_f, μ_f, σ_f) be a target triple. Assume that the $N \times N$ matrix $(\rho_{i,j})$ and the vector $(\tau_1, \cdots, \tau_N)$ are chosen so that there exists a non-constant solution $(\vec{S}, \vec{I})(t)$ of Model (4.6) and that its incidence has a finite mean time and a finite standard deviation. Then there exist constants $A, \phi > 0$ and a solution $(\vec{S_1}, \vec{I_1})(t)$ of Model (4.6; $(\rho_{j,k})$; $\phi\tau_1, \cdots, \phi\tau_N$) such that $g(t) := -A \sum_i p_i S_i'$ is a 3-statistic fit to (A_f, μ_f, σ_f).

We conjecture that for every non-constant solution $(\vec{S}, \vec{I})(t)$, the incidence $g(t)$ has a finite mean time and a finite standard deviation.

Proof. Let μ_0 and σ_0 denote the mean and standard deviation of the solution $(\vec{S}, \vec{I})(t)$. Let $\phi = \frac{\sigma_f}{\sigma_0}$. $(\vec{S}_\phi(t), \vec{I}_\phi(t)) = (\vec{S}(\frac{t}{\phi}), \vec{I}(\frac{t}{\phi}))$ is a solution to Model (4.6; $(\rho_{j,k})$; $\phi\tau_1, \cdots, \phi\tau_N$) such that $-\sum_i p_i S'_{\phi i}$ has a mean time of $\mu_0\phi$ and a standard deviation of $\sigma_0\phi = \sigma_f$.

Let $\lambda = \mu_f - \mu_0\ \phi$. $(\vec{S}^*_\lambda(t), \vec{I}^*_\lambda(t)) = (\vec{S}_\phi(t-\lambda), \vec{I}_\phi(t-\lambda))$ is a solution to Model (4.6; $(\rho_{j,k})$; $\phi\tau_1, \cdots, \phi\tau_N$) such that $-\sum_i p_i S^{*'}_{\lambda i}$ has a mean time of μ_f and a standard deviation of σ_f. Let $A = \frac{A_f}{-\sum_i \int p_i S^{*'}_{\lambda i} dt}$. Then the adjusted incidence $-A \sum_i p_i S^{*'}_{\lambda i}$ is A_f and that completes the proof. □

5. Relative least squares fits

In this section, using least squares fitting, or more precisely *relative* least squares fitting, we numerically illustrate that when real incidence data (New York City or London data) are used, the transmission parameters are practically unidentifiable. First, we state the definition of relative error; in our least squares fitting, we optimize the parameters with respect to this relative error.

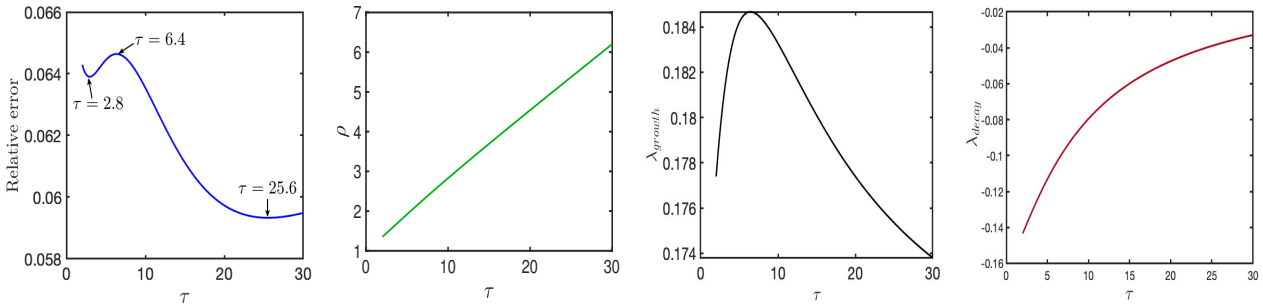
Definition 3 (relative error)**.** Suppose that the discrete sequences x_n and y_n are two representations of the daily new cases on day n of an outbreak; one might be actual data of new cases and the other a case history obtained from a mathematical model. We need a formula that tells us how similar y is to x. We say that the fit incidence y differs from the target incidence x by the **relative error (RE)**,

$$\mathrm{RE}(x(\cdot), y(\cdot)) \stackrel{\text{def}}{=\!=} \frac{\|x-y\|}{\|x\|} = \left[\frac{\sum_n [x_n - y_n]^2}{\sum_n x_n^2}\right]^{\frac{1}{2}}, \tag{5.1}$$

where $\|\cdot\|$ denotes the Euclidean norm. For continuous time, we use the integral version of Equation (5.1).

In this paper, the target $x(t)$ represents either the daily new case rate of an outbreak or an SIR solution. The fit y is always $-AS'(t)$, where $S(t)$ is a solution of Model (2.2) for some choice of parameters. We have four parameters ρ, τ, I_{t_0}, and A. We use the MATLAB built-in function **fmincon** for optimization. We use the default optimization algorithm, which uses an "**interior-point**" approach. This algorithm can be used to solve both linear and nonlinear optimization problems. We also use constrained optimization by setting the lower and upper bounds for our parameters. The lower bounds are set to ensure that the estimated parameter values are non-negative, as all our parameters are non-negative. The algorithm minimizes the relative error between a target data set and solutions to Model (2.2).

We use the NYC Omicron outbreak data (see Supplementary Material, Section B) as a target data set. For each value of τ in the interval (2.9, 25.6), we fix the value of τ and then optimize over the rest of the parameters: ρ, I_{t_0}, and A. The relative error corresponding to each value of τ is presented in panel (a) of Figure 6. For all values of τ in the interval (2.9, 25.6), the SIR least squares fits differ from the data by about the same amount, approximately 6%, see Figure 6(a). To illustrate that these fits are very close to each other and all are equally good fits to NYC data, we have selected the two sets of parameters that correspond to the two local minima in panel (a) of Figure 6, and we have sketched the corresponding outbreaks from Model (2.2) in Figure 2. The exact values of the parameters corresponding to these two least squares fits are shown in Table 1. The parameter values corresponding to each fit are significantly different from one another. For example, the $\tau = 2.9$ for the black curve and $\tau = 25.6$ for the red curve. So it is impossible to correctly estimate the outbreak characteristics using these fits. In fact, for any point chosen in the parameter space along the line shown in panel (b) of Figure 6, we obtain an equally good fit (in terms of relative error) to NYC data. This illustrates that ρ and τ are practically unidentifiable.



(a) **(b)** **(c)** **(d)**

Figure 6. How least squares fit to the NYC data varies as τ varies. For each τ value between 2 and 30, we find the least squares fit to NYC data by optimizing over the parameters ρ, A, and I_{t_0}. (a): Over a wide range of τ, the relative errors differ by less than 0.006, a factor of less than 10%. (b): For each τ, the corresponding ρ is shown. There is almost a linear relationship between τ and ρ, $0.2 \approx \frac{\rho-1}{\tau} = \lambda_{\text{growth}}$; see Eq. (3.4). (c): λ_{growth} varies by less than 6%. (d): λ_{decay} varies by a large factor.

The next real incidence data that we have used is the daily new cases of London Omicron data, see Supplementary Material, Section B. This data set has a dip in late December and early January. We believe this is a reporting problem due to the Christmas through New Year period, see Figure A.1 in the Supplementary Materials. To reflect this aberration, when we compute the least squares fits, we have excluded from consideration the interval from day 37 to day 50, that is, a fortnight surrounding the holiday period. Similar to the method we applied to NYC data, for each value of τ in the interval (0,30), we fix the value of τ and optimize the relative error over the rest of the parameters. The corresponding relative error as a function of τ is illustrated in Figure A.2. The two local least squares fits to London data are presented with solid curves in Figure 3. The estimated parameter values corresponding to these two least squares fits are given in Table 2. Even though these two fits both provide equally good fits, their corresponding estimated parameters are very different, meaning that it is not possible to uniquely identify the transmission parameters using limited, noisy data.

5.1. Nearly identical fits to simulated data are distinguishable by λ_{decay}, but they almost have the same λ_{growth}

In this section, we numerically illustrate that a wide range of parameter sets can provide very similar least squares fits to a simulated data set; however, those fits are not mathematically identical. In fact they are distinguishable by their tails.

As the first simulated data, we run Model (2.2) with $\tau = 7$, $\rho = 2$, $S_0 = 0.998$, and $I_0 = 0.001$, and store $-S'(t)$ as the simulated target data; $-S'(t)$ is the red curve in Figure 7. Then we fix the value of τ and optimize the relative error over the rest of the parameters, ρ, I_{t_0}, and A. Once the optimization function suggests a value for I_{t_0}, the value of S_{t_0} is calculated using Equation (2.4). We have repeated the process for the three different values of τ. The resulting least squares fits are shown in Figure 7. See Table 4 for the selected and the estimated parameter values corresponding to least squares fits provided in Figure 7. Figure 7 reveals that SIR solutions can be quite similar and differ significantly only in their final decay rates. Hence λ_{decay} is a reliable metric for describing SIR outbreaks.

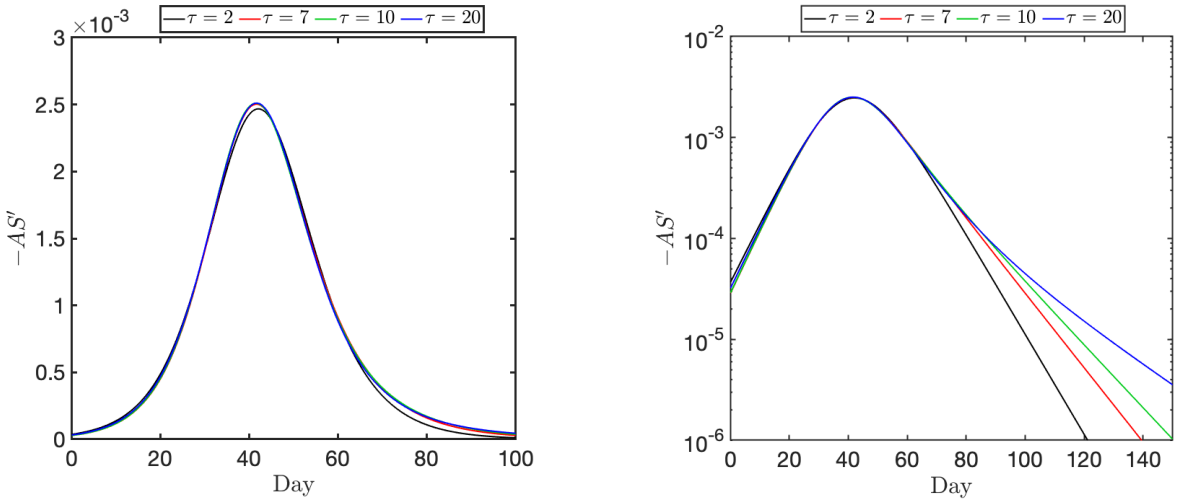


Figure 7. Least squares fits of SIR solutions to another SIR solution reveal a sensitive tail. The simulated data is obtained by solving Model (2.2) for $\rho = 2$ and $\tau = 7$. Then we find three least squares fits to that solution using $\tau = 2, 10$, and 20. For each optimization, we find the best values for the parameters ρ, I_{t_0}, and A such that the relative error is minimized; see Table 4 for the values of parameters in each case. Both panels show the same solutions. These plots show that the four solutions agree well with each other except in the decaying tail late in the outbreak, where the differences can only be seen clearly in the log scale used in the right panel.

Table 4. Parameter values corresponding to the four solutions in Figure 7 . The target has $\rho = 2$ and $\tau = 7$.

τ	ρ	A	I_0	S_0	Relative error	S_∞	λ_{growth}	λ_{decay}
2	1.2693	2.004	0.0003	0.9986	0.0306	0.608	0.1347	-0.1141
7	2	1.0	0.001	0.9980	–	0.2032	0.1429	-0.0848
10	2.4285	0.9082	0.0013	0.9978	0.007	0.1172	0.1429	-0.0715
20	3.7818	0.8269	0.0021	0.9972	0.0117	0.025	0.1391	-0.0453

To assess how λ_{growth} varies between different fits to a simulated data set, we conduct the following simulation: we choose four different values for $\rho = 1.5, 2, 3, 5$ and set the rest of the parameters as $\tau = 7$, $S_0 = 0.998$, and $I_0 = 0.001$. For each case, we save the corresponding $-S'(t)$ as the simulated data. We obtain four different simulated target data sets. For each target data and for each $\tau \in [3, 30]$, we optimize the relative error on the rest of the parameters I_{t_0}, A, and ρ. The selected and estimated parameters, for different simulated target data, are illustrated in Figure 8. For each case, the relationship between ρ and τ is approximately linear with the slope λ_{growth}. That means that for each simulated data, a family of parameter sets provides close least squares fits (see panel (a) for the relative error), and for each simulated data, all fits have almost the same exponential growth rate.

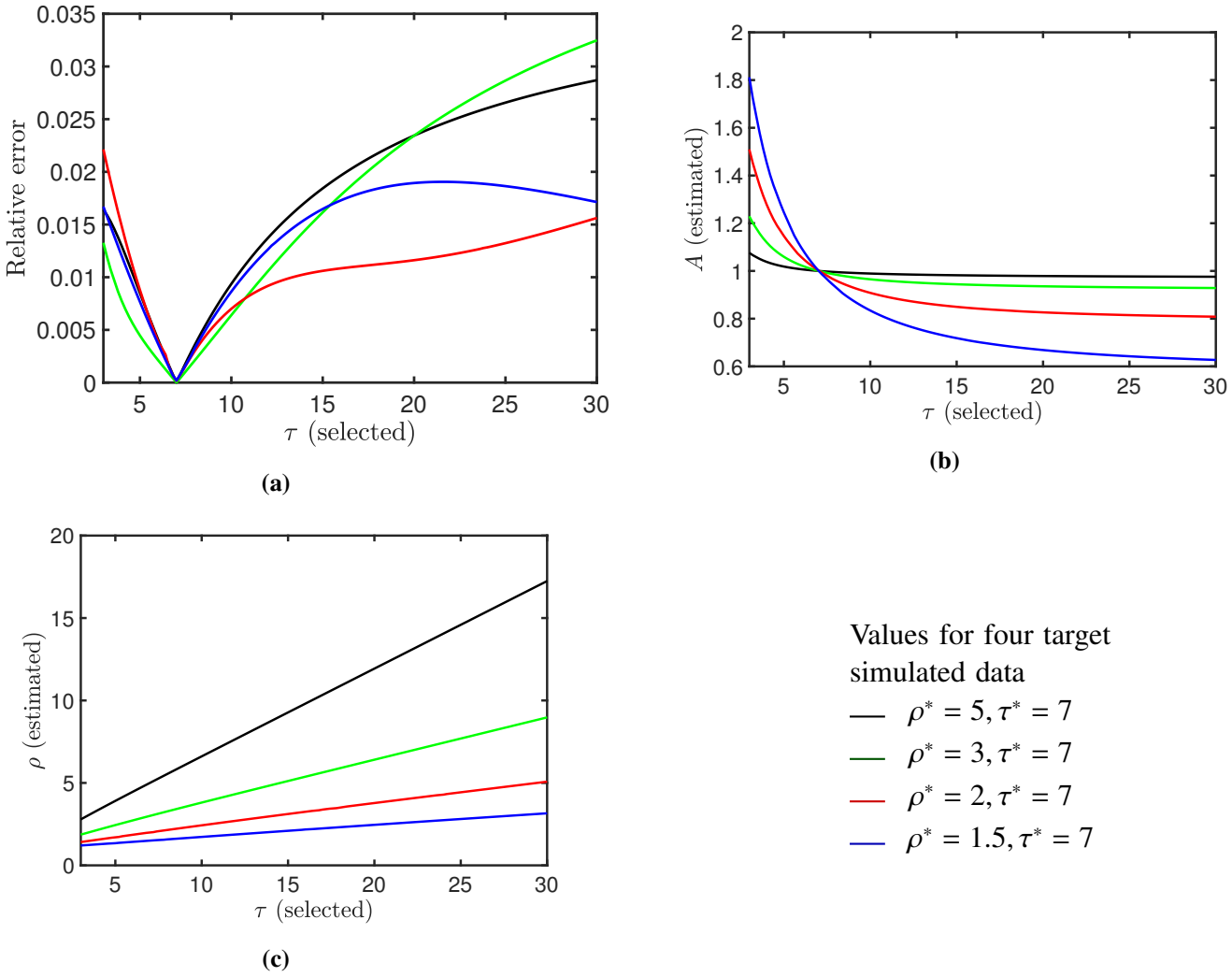


Figure 8. For each of four SIR simulated outbreaks, a family of parameters from least squares fits. We choose four solutions of Model (2.2) all with $\tau = 7$ and the ρ values $1.5, 2, 3, 5$. For each, we use its incidence $-S'(t)$ as a target data set. Using least squares, we fit the incidence rate of solutions of Model (2.2) to the four different target data. For each of the four target data sets, and for each value of τ from 3 to 30, we determine the least squares best fit set of parameters ρ, A, and I_{t_0} for which the incidence of the Model (2.2) solution has the smallest relative error. We show ρ (panel (a)) and A (panel (b)) as function of τ. In panel (a), there is almost a linear relationship between τ and ρ, where the slopes of the lines are approximately $\frac{\rho-1}{\tau} = \lambda_{\text{growth}}$; see Eq. (3.4) or Fig. 6(b). The four Model (2.2) targets $-S'$ have the same $\tau = 7$ and different ρ values as shown. They are initialized by setting $1 - S(0) = 0.001$, and then $I(0)$ is calculated from Equation (2.4), which implies that $S(-\infty) = 1$.

6. Discussion

One of the central challenges in the mathematical modeling of physical and biological systems is the identification of model parameters. It is well known that parameters of complex models often cannot be uniquely identified [40,41]. In this study, we consider a simple SIR model (Model (2.2)) and show theoretically that even such a basic model can exhibit parameter unidentifiability. Specifically, we investigate the identifiability of key transmission parameters (the basic reproduction number and the duration of infectiousness) in Model (2.2) from either incidence or prevalence data.

Recent studies in identifiability have shown that whether model parameters can be recovered depends strongly on the data type (e.g., prevalence versus incidence, early-phase versus full trajectory) [28, 29, 34]. Here, our results are consistent with these earlier finding. First, in Section 3, we show that when the incidence data ($-S'(t)$) from an SIR solution are used as the output, the basic reproduction number and the duration of infectiousness are structurally identifiable (see Theorem 1). Second, in Section 4, we show that if the three summary statistics (the mean, standard deviation, and total outbreak size) are used as the output, the model parameters are not identifiable (see Theorem 2).

Previous work has developed methods for determining whether the parameters of an input–output system are structurally globally identifiable. For example, Evans et al. [28] showed that when the output is prevalence data, the basic reproduction number for a version of the SIR model is structurally identifiable. Here we developed a method directly based on the properties of the SIR model and showed that the basic reproduction number and the duration of infectiousness can be uniquely calculated if the early exponential growth rate λ_{growth} and the late exponential decay rate λ_{decay} of an outbreak are known (see Theorem 1 and Proposition 1). This method remains primarily theoretical, since in practice the late exponential decay is often unavailable until the entire course of an epidemic is observed. As an illustration of how ρ and τ can be uniquely estimated using *Rat*, we applied the method to daily Omicron case data from New York City and London (see Figure 5 and Table 3).

Furthermore, we proved that the basic reproduction number and the duration of infectiousness are not practically identifiable from a target triple or three main statistics of an outbreak; that means if the mean time, standard deviation, and the size of an outbreak are given, it is not possible to uniquely identify the key transmission parameters. In fact, we showed that for each choice of the basic reproduction number, one could find a 3-statistic fit (see Theorem 2). Hence, it is impossible to find a good estimation for the basic reproduction number. We also extend our result to an SIR model with N compartments. We show that an N-compartment SIR model admits an N^2-dimensional family of parameter matrices that can be adjusted (via time-scaling and shifts) to produce identical 3-statistic fits (see Theorem 3). That means that multi-compartmental models admit vastly larger families of indistinguishable parameter sets.

We also numerically illustrated that when real incidence data, such as those from New York City or London, are used, the transmission parameters are practically unidentifiable. Our simulations showed that distinct parameter vectors yield nearly identical fits; see Figures 2 and 3. In this case, the only parameter that can be uniquely identified to two decimal places is λ_{growth}; see Table 1 and Table 2.

Overall, this work shows that even simple epidemic models can conceal deep ambiguities in their parameters. Identifiability is not guaranteed by model simplicity but by the nature and completeness of the data used to constrain it. Recognizing this limitation is essential for interpreting fitted epidemic parameters and for designing data collection strategies that can make such parameters truly measurable.

By quantifying when and why identifiability fails, our study helps set a foundation for more transparent and reliable use of mechanistic models in epidemiology.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. H. W. Hethcote, The mathematics of infectious diseases, *SIAM Review*, **42** (2000), 599–653. https://doi.org/10.1137/S0036144500371907
2. F. Brauer, P. Van den Driessche, J. Wu, L. J. S. Allen, *Mathematical epidemiology*, volume 1945. Springer, 2008. https://doi.org/10.1007/978-3-540-78911-6
3. F. Brauer, C. Castillo-Chavez, Z. Feng, *Mathematical models in epidemiology*, volume 32. Springer, 2019. https://doi.org/10.1007/978-1-4939-9828-9
4. R. Ross, *The prevention of malaria*, John Murray, 1911.
5. W. O. Kermack, A. G. McKendrick, A contribution to the mathematical theory of epidemics, *Proceedings of the Royal Society A*, **115** (1927), 700–721. https://doi.org/10.1098/rspa.1927.0118
6. Y. Li, Y. Yao, M. Feng, T. P. Benko, M. Perc, J. Zavrsnik, Epidemic dynamics in homes and destinations under recurrent mobility patterns, *Chaos Solit. Fract.*, **195** (2025), 116273. https://doi.org/10.1016/j.chaos.2025.116273
7. J. Chen, C. Xia, M. Perc. The siqrs propagation model with quarantine on simplicial complexes, *IEEE Transactions on Computational Social Systems*, **11** (2024), 4267–4278. https://doi.org/10.1109/TCSS.2024.3351173
8. W. P. London, J. A. Yorke. Recurrent outbreaks of measles, chicken pox, and mumps: I. Seasonal variation in contact rates, *Am. J. Epidemiol.*, **98** (1973), 453–468. https://doi.org/10.1093/oxfordjournals.aje.a121575
9. A. S. Mahmud, C. J. E. Metcalf, B. T. Grenfell, Comparative dynamics, seasonality in transmission, and predictability of childhood infections in Mexico, *Epidemiol. Infect.*, **145** (2017), 607–625. https://doi.org/10.1017/S0950268816002673
10. A. Rachah, D. F. M. Torres, Modeling, dynamics and optimal control of Ebola transmission. *Mathematics in Computer Science*, **10** (2016), 331–342. https://doi.org/10.1007/s11786-016-0268-y
11. F. Oduro, G. Appaboah, J. Baafi, Optimal control of Ebola transmission dynamics with interventions, *British J. Math. Computer Sci.*, **19** (2016), 1–19.
12. R. Kumar, S. Dey, Sir model for Ebola outbreak in Liberia, *Int. J. Math. Trends Technol.*, **28** (2015), 28–30.

13. I. E. Kibona, C. Yang, SIR Model of spread of Zika virus infections: ZikV linked to microcephaly simulations, *Health*, **9** (2017), 1190–1210.

14. S. E. B. Boret, R. Escalante, M. Villasana, Mathematical modeling of Zika virus in Brazil, (2023), *arXiv preprint*, arXiv: 1708.01280v2.

15. R. Dohare, M. Kumar, S. Sankhwar, N. Kumar, S. K Sagar, J. Kishore, SIR-SI model for Zika virus progression dynamics in India: A Case study, *J. Commun. Diseases*, **53** (2021), 100–104.

16. A. D. Zewdie, S. Gakkhar, A Mathematical model for Nipah virus infection. *J. Appl. Math.*, **2020** (2020), (6050834). https://doi.org/10.1155/2020/6050834

17. A. K. Sikdar, B. Hossain, H. Islam, Compartmental modeling in epidemic diseases: Comparison between SIR model with constant and time-dependent parameters, *Inverse Problems*, **39** (2023), 035055. https://doi.org/10.1088/1361-6420/acb4e7

18. K. R. Kumar, Nipah outbreak in Kerala: A network-based study, *J. Phys. Conference Ser.*, **1850** (2000), 012019. https://doi.org/10.1088/1742-6596/1850/1/012019

19. I. Cooper, A. Mondal, C. G. Antonopoulos, A SIR model assumption for the spread of COVID-19 in different communities, *Chaos Solit. Fract.*, **139** (2020), 110057. https://doi.org/10.1016/j.chaos.2020.110057

20. N. A. Kudryashov, M. A. Chmykhov, M. Vigdorowitsch, Analytical features of the SIR model and their applications to COVID-19, *Appl. Math. Model.*, **90** (2021), 466–473. https://doi.org/10.1016/j.apm.2020.08.057

21. J. David, et. al. *Mathematical Models: Perspectives of Mathematical Modelers and Public Health Professionals*, 2023.

22. O. Diekmann, J. A. P. Heesterbeek, J. A. J. Metz, On the definition and the computation of the basic reproduction ratio r 0 in models for infectious diseases in heterogeneous populations, *J. Math. Biol.*, **28** (1990), 365–382. https://doi.org/10.1007/BF00178324

23. P. van den Driessche, J. Watmough, Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission, *Math. Biosci.*, **180** (2002), 29–48. https://doi.org/10.1016/S0025-5564(02)00108-6

24. B. Dhungel, Md. S. Rahman, Md. M. Rahman, A. K. C. Bhandari, P. M. Le, N. A. Biva, et al., Reliability of early estimates of the basic reproduction number of covid-19: A systematic review and meta-analysis, *Int. J. Environ. Res. Public Health*, **19** (2022), 11613. https://doi.org/10.3390/ijerph191811613

25. S. Jahedi, J. A. Yorke, When the best pandemic models are the simplest, *MDPI Biology*, **9** (2020), 353. https://doi.org/10.3390/biology9110353

26. A. F. Villaverde, N. Tsiantis, J. R. Banga, Full observability and estimation of unknown inputs, states and parameters of nonlinear biological models, *J. Royal Soc. Interf.*, **16** (2019), 20190043. https://doi.org/10.1098/rsif.2019.0043

27. R. Bellman, K. J. Astrom, On structural identifiability, *Math. Biosci.*, **7** (1970), 329–339. https://doi.org/10.1016/0025-5564(70)90132-X

28. 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. https://doi.org/10.1016/j.mbs.2004.10.011

29. N. Tuncer, T. T. Le, Structural and practical identifiability analysis of outbreak models, *Math. Biosci.*, **299** (2018), 1–18. https://doi.org/10.1016/j.mbs.2018.02.004 .

30. N. Meshkat, A. Shiu, Structural identifiability of compartmental models: Recent progress and future directions, *arXiv preprint arXiv:2507.04496*, 2025.

31. E. A. Dankwa, A. F. Brouwer, C. A. Donnelly, Structural identifiability of compartmental models for infectious disease transmission is influenced by data type, *Epidemics*, **41** (2022), 100643. https://doi.org/10.1016/j.epidem.2022.100643

32. W. C. Roda, M. B. Varughese, D. Han, M. Y. Li, Why is it difficult to accurately predict the COVID-19 epidemic? *Inf. Dis. Modelling*, **5** (2020), 271–281. https://doi.org/10.1016/j.idm.2020.03.001

33. F. Bergström, M. Favero, T. Britton, Identifiability in epidemic models with prior immunity and under-reporting, 2025. URL `http://arxiv.org/abs/2506.07825`

34. T. Sauer, T. Berry, D. Ebeigbe, M. M. Norton, A. J. Whalen, S. J. Schiff, Identifiability of infection model parameters early in an epidemic. *SIAM J. Control and Optimization*, **60** (2022), S27–S48. https://doi.org/10.1137/20M1353289

35. E. D. Sontag, Dynamic compensation, parameter identifiability, and equivariances, *PLOS Comp. Bio.*, **13** (2017), e1005447. https://doi.org/10.1371/journal.pcbi.1005447

36. J. C. Miller, A note on the derivation of epidemic final sizes, *Bull.Math. Bio.*,**74** (2012), 2125–2141. https://doi.org/10.1007/s11538-012-9749-6

37. M. Turkyilmazoglu, Explicit formulae for the peak time of an epidemic from the SIR model, *Phys. D Nonlinear Phenom.*, **422** (2021), 132902. https://doi.org/10.1016/j.physd.2021.132902

38. F. Brauer, Early estimates of epidemic final sizes, *J. Bio. Dyn.*, **13** (2019), 23–30. https://doi.org/10.1080/17513758.2018.1469792

39. C. Meyer, *Matrix analysis and applied linear algebra*, Texts in Applied Mathematics. Siam, Philadelphia, USA, 2000. ISBN 9781489976123.

40. R. C. Spear, Large simulation models: calibration, uniqueness and goodness of fit, *Env. Modelling Software*, **12** (1997), 219–228. https://doi.org/10.1016/S1364-8152(97)00014-5

41. M. K. Transtrum, B. B. Machta, J. P. Sethna, Why are nonlinear fits to data so challenging?, *Phys. Rev. Lett.*, **104** (2010), 060201. https://doi.org/10.1103/PhysRevLett.104.060201

Supplementary material

A. Figures

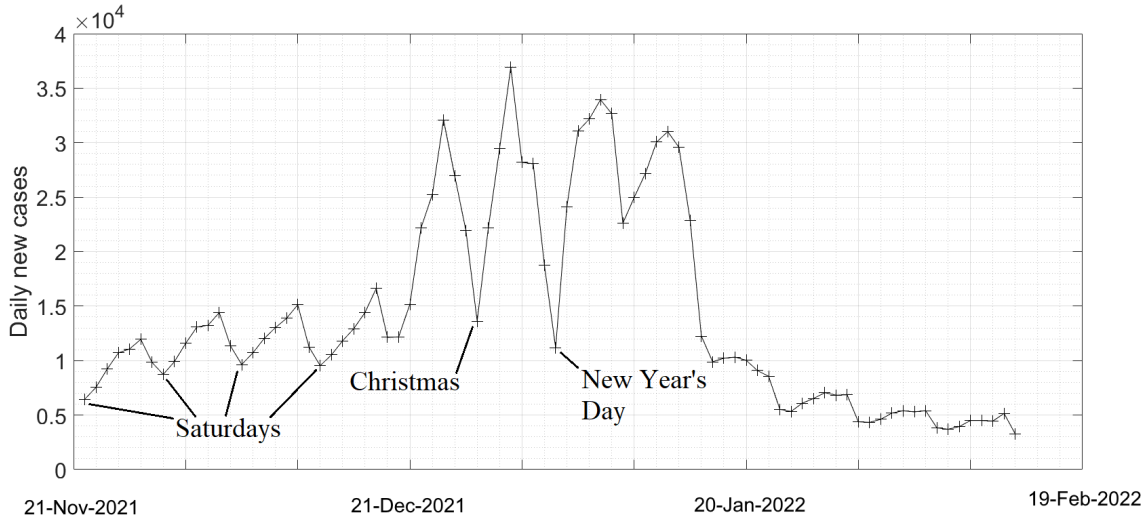


Figure A.1. The daily raw data of all COVID-19 cases for London.

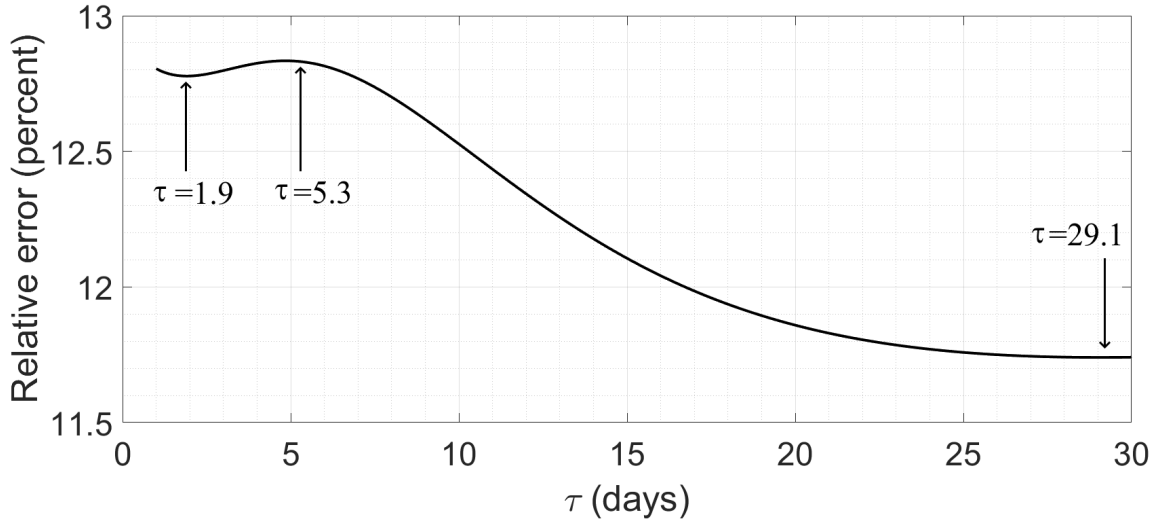


Figure A.2. Fit of the London data using SIR model with the central period excluded. The relative error has two local minima at $\tau = 1.9$ and $\tau = 29.1$ days. Further details of these fits are in Table 2.

B. Data

B.1. COVID-19 Omicron Data

B.1.1. New York City

We harvested the raw daily case data from
https://www.nyc.gov/site/doh/covid/covid-19-data-totals.page. This page provides a csv file from which we used the seven-day moving (centered) average of confirmed daily case counts (Column D of the csv). We selected the period from November 28, 2021 to February 19, 2022 as the time interval of interest for Omicron, amounting to 84 days in total. The starting date was when daily cases of B1.1.529 reached 30, and the end date is when the B1.1.529 epidemic was decreasing and being replaced by BA.2 and its sublineages.

The NYC local website does not provide data related to cases by variant. Hence, we used the CDC data from https://covid.cdc.gov/covid-data-tracker/#datatracker-home for this. The fraction of cases for each variant is reported weekly, and by geographical regions, which are composed of several states combined together. Region 2 consists of New York, New Jersey, Puerto Rico, and the Virgin Islands. We report these proportions during the period in question. To fit the continuous models, i.e., differential equation models, we converted the weekly data reporting the fraction for each strain to daily data using the "makima" interpolation routine in Matlab. Finally, we multiplied the daily proportion of cases (for Omicron) by the daily case reports of all COVID cases to obtain the presumed daily counts of the Omicron variant in NYC. The dots in Figure 2 represent the data after applying a moving seven-day average that eliminates within-week regular fluctuations.

NYC total population size: 1,024,164

The variant proportion of Omicron in US HHS Region 2 over the twelve-week period under consideration is as follows :

2.1 19.2 55.7 84.7 90.1 95.7 97.9 98.7 97.9 98.2

NYC's moving (centered) seven-day average of daily cases of the Omicron variant for 84 consecutive days beginning Nov. 28, 2021. The cases of each date are averaged with the three previous days and the following three days. This data is shown with ten consecutive days per row:

32	53	87	135	211	290	375	470	589	755
955	1178	1442	1673	1968	2597	3666	5083	6747	8447
9855	11158	14036	16818	19645	22252	22963	22814	24763	28328
31161	34062	36696	38696	39641	41217	41945	42267	40814	39136
38320	39005	36322	31771	27050	23724	20648	19137	17651	16448
14047	12861	11354	9932	8737	8250	7708	7102	6140	5391
4873	4398	3900	3542	3263	2850	2490	2222	1987	1816
1869	1785	1570	1462	1317	1203	1119	1069	1005	918
832	772	625	659						

B.1.2. London

The raw data was harvested from
https://coronavirus.data.gov.uk/details/cases?areaType=region&areaName=London.

We selected the period from November 21, 2021 to February 12, 2022 as the duration to focus on, amounting to 84 days in total. In this case, the data itself does not include seven-day averaging, so we implemented it manually. The variant proportions are available for the UK as a whole and on a weekly basis. Once again, we used the interpolation routine to convert it to a daily proportion and thereby obtained a daily case history.

London total population size : 906,656

The variant proportion of Omicron in the UK over the twelve-week period under consideration is as follows :

0.2 1.4 13.2 58.0 82.3 94.6 94.1 97.4 96.2 92.8 85.4 76.5

London's moving seven-day average of daily cases of the Omicron variant for 84 consecutive days beginning Nov. 21, 2021. The red points in the London graph and the red points in the table below represent smooth data from the winter vacation period. These were not used in fitting the curve to data in Figure 3. This data is shown with ten consecutive days per row:

17	23	36	53	74	98	125	152	209	324
493	709	970	1268	1591	2089	2833	3709	4653	5575
6397	7047	7623	8269	8838	9563	10492	12060	13868	16374
18086	19776	20400	21705	22977	24750	24351	24626	24306	24067
24410	24694	24067	24884	25545	26101	28034	28471	28331	28162
27532	26132	24654	22522	20133	17371	14445	11599	9604	8645
7983	7382	6826	6384	6036	5754	5543	5346	5089	4854
4588	4338	4104	3978	3843	3706	3567	3407	3256	3168
3049	2942	2858	2743						