Mathematics

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

AIMS Mathematics, 9(9): 24722–24761.
DOI: 10.3934/math.20241204
Received: 27 April 2024
Revised: 16 July 2024
Accepted: 12 August 2024
Published: 23 August 2024

Research article

Comparative analysis of practical identifiability methods for an SEIR model

Omar Saucedo[1,*], Amanda Laubmeier[2], Tingting Tang[3], Benjamin Levy[4], Lale Asik[5], Tim Pollington[6] and Olivia Prosper Feldman[7]

[1] Department of Mathematics, Virginia Tech, Blacksburg, VA 24061, USA

[2] Department of Mathematics and Statistics, Texas Tech University, Lubbock, TX 79409, USA

[3] Department of Mathematics and Statistics, San Diego State University, San Diego, CA 92182, USA

[4] Division of Mathematics, Analytics, Science, and Technology, Babson College, Wellesley, MA 02481, USA

[5] Department of Mathematics and Statistics, University of the Incarnate Word, San Antonio TX 78209, USA

[6] Big Data Institute, University of Oxford, OX1 2JD, Oxford, UK

[7] Department of Mathematics, University of Tennessee, Knoxville, TN 37996, USA

* **Correspondence:** Email: osaucedo@vt.edu; Tel: +1-540-231-8283.

Abstract: Identifiability of a mathematical model plays a crucial role in the parameterization of the model. In this study, we established the structural identifiability of a susceptible-exposed-infected-recovered (SEIR) model given different combinations of input data and investigated practical identifiability with respect to different observable data, data frequency, and noise distributions. The practical identifiability was explored by both Monte Carlo simulations and a correlation matrix approach. Our results showed that practical identifiability benefits from higher data frequency and data from the peak of an outbreak. The incidence data gave the best practical identifiability results compared to prevalence and cumulative data. In addition, we compared and distinguished the practical identifiability by Monte Carlo simulations and a correlation matrix approach, providing insights into when to use which method for other applications.

Keywords: identifiability; parameter estimation; Monte Carlo; correlation matrix; epidemiological data; mathematical modeling
Mathematics Subject Classification: 92-10, 92D30, 34C60

1. Introduction

Compartment models have been used extensively to study infectious diseases. Among them, susceptible-exposed-infectious-recovered (SEIR) models have been extensively used to study disease dynamics impacted by vertical transmission [1, 2], vaccination strategies [3], delayed infectiousness [4], multistage infectiousness [5], and treatment strategies [6], and most recently, were applied to COVID-19 [7–9]. These models often include parameters for which numerical values are unknown *a priori* and cannot be directly measured. Since many parameters in a given model are not directly measurable, researchers often obtain parameters from outbreak data.

Parameter estimation relies on comparing empirical observations of the modeled system with the corresponding model output. Many computational techniques can be employed for parameter estimation. Most of these techniques rely on minimizing the difference between model output and observed data. However, before numerically estimating parameter values, it is important to address whether the parameters of the model are identifiable. Identifiability analysis determines to what extent and with what level of certainty the parameters of a model can be recovered from the available empirical data [10]. These relate to two general types of identifiability analysis: structural and practical identifiability. Structural identifiability is the theoretical possibility of determining the true values of parameters of a model from observations of its outputs and knowledge of its dynamic equations [11–13]. On the other hand, practical identifiability provides information on the accuracy with which parameters can be estimated from the available discrete and noise-corrupted measurements [14]. So, for an infinite amount of noise-free data, structural identifiability implies (the maximum) practical identifiability [15].

Three common methods for evaluating the practical identifiability of models are Monte Carlo simulations [16, 17], correlation matrices [18–20], and the profile likelihood [21–25]. The Monte Carlo approach can be applied to evaluate disparate model structures each with different observation schema by implementing a random sampling algorithm [16]. While Monte Carlo simulations offer conceptual and algorithmic simplicity, their computational cost can be very high, as many samples are required to obtain a good approximation. The Correlation Matrix approach assesses the correlation between parameter estimates [18–20]. It is much less computationally intensive than the Monte Carlo method but only provides a pairwise analysis of model parameters. Likelihood profiling is a common way to perform a practical identifiability analysis when using a likelihood-based estimation procedure (e.g., approximate Bayesian computation or maximum likelihood estimation) because the profiles are used for quantifying uncertainty in parameter estimates (i.e., to approximate confidence intervals) [23].

Epidemiological data plays a critical role in infectious disease surveillance, as it describes how infectious diseases are distributed within populations and what factors contribute to transmission [26]. Several mathematical models have been established during the last decades to forecast disease progression using epidemiological data [27–32]. There are many technical challenges in implementing a standardized epidemiological data framework. The way data is reported and collected, and the type and frequency of data available through data-sharing institutions (public health departments, ministries of health, data collection, or aggregation services) differ significantly. For example, different data types (including prevalence, incidence, cumulative incidence, and mortality) are reported at differing time intervals (such as daily, weekly, monthly, or even more infrequently). As a result, understanding how these factors affect parameter identifiability and estimation is critical.

In this study, we systematically examined the influence of different data types (prevalence, incidence, and cumulative incidence) and sampling frequencies (daily, weekly, and monthly) on the practical identifiability of SEIR model parameters. We explored four scenarios with different peak infection times and data collection windows to assess the impact of different data collection strategies on parameter identifiability. Discrepancies between Monte Carlo and correlation matrix results highlight the importance of considering multiple criteria for identifiability. Overall, our results show that incidence data yielded the most identifiable parameters, prevalence data exhibited intermediate identifiability, and cumulative incidence data resulted in the least identifiable parameters. Varying data collection frequencies affected parameter identifiability, with longer time series and higher sampling rates generally improving identifiability.

In Section 2, we present the SEIR model and the structural identifiability results. Sections 3.1 and 3.2 outline the use of MC and CM methods to assess parameter identifiability, respectively. In Section 4, we present the results of our analysis, with Section 4.1 detailing the outcomes of the MC simulations, Section 4.2 illustrating the CM results, and Section 4.3 comparing the results from both MC and CM methods. In Section 5, we discuss our finding, their implications, and limitations. Finally, in Section 6, we provide a concise summary of our findings.

2. The SEIR model

The overarching goal of our study is to examine how different data types and collection frequencies influence parameter identifiability under structural and practical identifiability paradigms. To this end, we chose a simple SEIR model framework containing a small number of parameters as the foundation for our study. Here, the SEIR model is a closed population with no vital dynamics (i.e., no births or deaths):

$$\dot{S}(t) = -\beta S I, \quad \dot{E}(t) = \beta S I - \gamma E, \quad \dot{I}(t) = \gamma E - \alpha I, \quad \dot{R}(t) = \alpha I, \quad \dot{C}(t) = \beta S I, \tag{2.1}$$

where $N(t) = S(t) + E(t) + I(t) + R(t)$ is the total population, the parameter β is the transmission rate from susceptible to infected, γ is the rate of transition from exposed to infectious, and α is the recovery rate from infected to recovered (Figure 1). The state variable $I(t)$ represents the prevalence of the infectious disease at time t. The auxiliary variable $C(t)$ tracks the cumulative number of infectious individuals at time t since the start of the outbreak. $C(t)$ is not a state of the system but rather a class to track cumulative incidence, from which incidence can be derived. Thus, the instantaneous number of new infections (i.e., incidence) at time t is $\dot{C}(t)$. In practice, the incidence of new cases is measured over a time interval $[t_{i-1}, t_i]$, and therefore can be computed using the difference of the cumulative incidence at these time points: $C(t_i) - C(t_{i-1})$. Initial conditions for the S, E, I, and R states will be noted by $S(0), E(0), I(0)$, and $R(0)$, respectively.

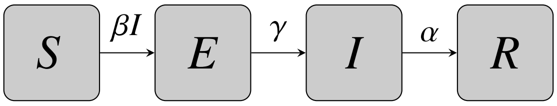


Figure 1. Flowchart for the general SEIR framework.

The SEIR model Eq (2.1) is well studied and often presented as an entry-level educational tool [33]. As individuals progress through the compartments, eventually the number of susceptible individuals may be insufficient to sustain transmission. This leads to the disease-free equilibrium, which is globally asymptotically stable when $\beta N/\alpha < 1$ and unstable when $\beta N/\alpha > 1$. This cutoff is the basic reproduction number $\mathcal{R}_0$, which is commonly used to characterize the infectiousness of a disease that is freshly introduced into a completely susceptible population. Importantly, this reproduction number $\mathcal{R}_0$, along with other epidemiological metrics, are calculated from model parameters. If a model is non-identifiable, then researchers risk reaching incorrect conclusions about disease persistence that follow from incorrect parameter values.

2.1. Structural identifiability

Structural identifiability addresses the question of whether the parameters of a model can be uniquely determined from "perfect" input-output data, assuming the underlying model correctly represents the input-output data. Here, *perfect data* means the data is continuous in time and without noise. Structural identifiability serves as a prerequisite for practical identifiability analyses as it sets the foundation on whether the model parameters can be distinctly ascertained. Without first establishing local or global identifiability, the results attained in practical identifiability analysis can be challenging to interpret due to uncertainty in the numerical optimization. Consider a general compartment model:

$$\begin{aligned}\dot{\mathbf{x}}(t) &= f(x(t), \mathbf{p}),\\ \mathbf{x}(0) &= \mathbf{x}_0,\\ \mathbf{y}(t) &= \mathbf{g}(\mathbf{x}(t), \mathbf{p}),\end{aligned} \tag{2.2}$$

where the n state variables at time t are represented by the state vector $\mathbf{x}(t) \in \mathbb{R}^n$, parameters are denoted by the vector $\mathbf{p} \in \mathbb{R}^k$, the initial condition vector is denoted as $\mathbf{x}_0 \in \mathbb{R}^n$, and the m observation variables are represented by $\mathbf{y}(t) \in \mathbb{R}^m$. The system Eq (2.2) is globally structurally identifiable for the parameter vector $\mathbf{p}_1$ if for every parameter vector $\mathbf{p}_2$,

$$y(t, \mathbf{p}_1) = y(t, \mathbf{p}_2) \implies \mathbf{p}_1 = \mathbf{p}_2.$$

We say that system Eq (2.2) is locally structurally identifiable for the parameter vector $\mathbf{p}_1$ if, for every $\mathbf{p}_2$,

$$y(t, \mathbf{p}_1) = y(t, \mathbf{p}_2), \text{ and } \mathbf{p}_2 \in B(\mathbf{p}_1) \implies \mathbf{p}_1 = \mathbf{p}_2, \text{ where } B(\mathbf{p}_1) \text{ is a ball centered at } \mathbf{p}_1.$$

Otherwise, system Eq (2.2) is unidentifiable. An equivalent characterization of local identifiability is if there are at least two or more finite parameter values of $\mathbf{p}$ that yield the same observations $\mathbf{y}$. Here, the state variables denoted by $\mathbf{x}$ are specifically $[x_1, x_2, x_3, x_4, x_5] = [S, E, I, R, C]$.

There are several well-established methods for determining the structural identifiability of disease models such as the direct test [34, 35] and the similarity transformation method [36–38], which can be applied to autonomous (no external input) systems. Alternatively, the Taylor series approach [39], differential algebra [40, 41], the generating series approach [42], implicit functions approach [43], or differential geometry [44] are applicable to external input. Various software tools have been developed to perform structural identifiability analysis, such as the observability test [45], differential algebra for identifiability of systems (DAISY, [40]), exact arithmetic rank (EAR, [46]), COMBOS [47], Data2Dynamics [48], STRIKE-GOLDD [44], GenSSI2 [49], structural identifiability analyser (SIAN, [50]), and StructualIdentifiablity.jl [51]. These toolboxes provide a broad range of features and implement methods based on differential algebra, semi-numerical differential algebra, generating series, local algebraic observability, and identifiability tables. However, some of these toolboxes have limited performance or can be time-consuming to use [52].

We establish the structural identifiability of the prevalence and cumulative incidence data types using the differential algebra approach with the aid of the software package `StructuralIdentifiability.jl.` The details are given in the following propositions.

Proposition 1. *The system Eq* (2.1) *is locally identifiable to prevalence data and globally identifiable if initial conditions are known.*

Proof. Given the observation is prevalence data $I(t)$, that is, $y(t) = I(t)$, the input-output equation derived by `StructuralIdentifiability.jl` is:

$$II''' - I'I'' + (\alpha + \gamma)II'' + \beta I^2 I'' - (\alpha + \gamma)(I')^2 + (\alpha + \gamma)\beta I^2 I' + \alpha\beta\gamma I^3 = 0.$$

In this equation, I' represents the derivative of the state variable $I(t)$. Using the definition of being globally identifiable, we set $y(t, \mathbf{p}_1) = y(t, \mathbf{p}_2)$, where $\mathbf{p}_1 = [\alpha_1, \beta_1, \gamma_1]$, and $\mathbf{p}_2 = [\alpha_2, \beta_2, \gamma_2]$ which yields the following equations:

$$\alpha_1 + \gamma_1 = \alpha_2 + \gamma_2, \quad (\alpha_1 + \gamma_1)\beta_1 = (\alpha_2 + \gamma_2)\beta_2, \quad \beta_1 = \beta_2, \quad \alpha_1\gamma_1\beta_1 = \alpha_2\gamma_2\beta_2.$$

Solving the system of equations above using the Groebner basis approach in terms of $\mathbf{p}_1$ in `Python`, we obtain two sets of solutions:

$$\{\alpha_1 = \gamma_2, \beta_1 = \beta_2, \gamma_1 = \alpha_2\} \qquad \text{and} \qquad \{\alpha_1 = \alpha_2, \beta_1 = \beta_2, \gamma_1 = \gamma_2\}.$$

Hence, the system Eq (2.1) is locally identifable given prevalence data. Furthermore, the function `find ioequations` within `StructuralIdentifiability.jl` outputs all identifiable equations including the initial conditions which are listed below:

$$I(t), \quad \beta, \quad S(t)\gamma, \quad \alpha\gamma, \quad \alpha + \gamma, \quad E(t)\gamma + I(t)\gamma.$$

It is clear to see that, with the initial condition $S(0), E(0),$ and $I(0)$ known, the identifiable equation $S(t)\gamma$ combined with $\alpha\gamma$ implies that both α and γ become globally identifiable. □

Proposition 2. *The system Eq* (2.1) *is locally identifiable to cumulative data and globally identifiable if initial conditions are known.*

Proof. Given the observation being the prevalence data $C(t)$, that is, $y(t) = C(t)$, we can obtain the input-output equation using `StructuralIdentifiability.jl`. Following a similar procedure as in Proposition 2.1, we obtain the following set of equations:

$$\alpha_1\gamma_1 = \alpha_2\gamma_2, \quad \beta_1\gamma_1 = \beta_2\gamma_2, \quad \beta_1 + \gamma_1 = \beta_2 + \gamma_2.$$

We solve this system using the Groebner basis approach in terms of $\mathbf{p}_1$ in `Python` and obtain two sets of solutions:

$$\{\alpha_1 = \alpha_2, \beta_1 = \beta_2, \gamma_1 = \gamma_2\} \qquad \text{and} \qquad \{\alpha_1 = \frac{\alpha_2\gamma_2}{\beta_2}, \beta_1 = \gamma_2, \gamma_1 = \beta_2\}.$$

Thus, the system Eq (2.1) is locally identifiable given cumulative data.

Furthermore, to investigate the impact of initial conditions, we can find all identifiable equations including the initial conditions:

$$C(t), \quad S(t), \quad I(t)\beta, \quad \beta\gamma, \quad I(t)\alpha, \quad E(t) + I(t), \quad \alpha + \gamma.$$

It is clear to see that, with the initial condition $S(0), E(0),$ and $I(0)$ known, the identifiable equation $I(t)\beta, \beta\gamma$ combined with $\alpha + \gamma$ implies all three parameters become globally identifiable. □

As far as we know, all of these packages can only handle observations when explicitly expressed as functions of the state variables at any time. Consequently, we are unable to examine the structural identifiability of the SEIR model for incidence defined as a difference in time, so its analysis will be limited to practical identifiability. Table 1 summarizes the outcomes of these assessments. In cases with specified initial conditions, all parameters of the SEIR model exhibit structural identifiability concerning both prevalence and cumulative incidence outputs. However, when initial conditions are unspecified, only β is structurally identifiable with respect to prevalence, while the remaining parameters are only locally identifiable. Cumulative incidence shows all parameters are only locally identifiable.

Table 1. Structural identifiability of $\beta, \gamma,$ and α with or without initial conditions (ICs) for different observables. The results for the cases without ICs using both SIAN and DAISY were identical.

Observables	Globally Identifiable	Locally Identifiable
Prevalence (I)	β	γ, α
Prevalence (I with ICs)	β, γ, α	
Cumulative (C)		β, γ, α
Cumulative (C with ICs)	β, γ, α	

3. Methods for practical identifiability

Compared to structural identifiability analysis, practical identifiability analysis accounts for the sampling rates and noisiness of experimental data [53]. For this reason, practical identifiability analyses typically involve fitting models to epidemic data. In general, a model is considered practically identifiable if a unique parameter vector for a given model can be consistently obtained. Researchers may use real or simulated epidemic data to assess practical identifiability [17, 54–56], and there are different approaches and criteria for assessing identifiability. In this study, we apply two approaches to analyze the practical identifiability of a SEIR model with hypothetical "true" parameters. We first use a Monte Carlo approach to consider the practical identifiability of our system of ODEs before conducting a similar analysis using a correlation matrix approach.

In general, practical identifiability can be defined for the dynamical model in Eq (2.2). In this study, we consider $\mathbf{x} = [S, E, I, R, C]$ and $f(\mathbf{x}(t), \mathbf{p})$ the right-hand side of Eq (2.1). We assume that $\mathbf{x}_0$ is known but that all parameters $\mathbf{p} = [\beta, \gamma, \alpha]$ must be recovered from data. For practical identifiability, observations $\mathbf{y}(t)$ at a finite set of time points t_i are modeled by:

$$\mathbf{y}(t_i) = g(\mathbf{x}(t_i), \mathbf{p})(1 + \epsilon(t_i)), \tag{3.1}$$

with random measurement noise $\epsilon(t_i)$ drawn from a normal distribution with zero mean and standard deviation σ, i.e., $\epsilon \sim \mathcal{N}(0, \sigma)$. The set of time points t_i are sampled at different frequencies and time spans depending on the scenario. In this study, we denote $\mathbf{y}$ as prevalence, incidence, or cumulative incidence data and define $g(\mathbf{x}(t), \mathbf{p})$ to select the appropriate state: $g(\mathbf{x}(t_i), \mathbf{p}) = x_3(t_i)$ for prevalence, $g(\mathbf{x}(t_i), \mathbf{p}) = x_5(t_i)$ for cumulative incidence, and $g(\mathbf{x}(t_i), \mathbf{p}) = x_5(t_i) - x_5(t_{i-1})$ for incidence. A model is practically identifiable if there is a unique parameter vector $\mathbf{p} = \hat{\mathbf{p}}$ that minimizes the difference between the model output g and the data $\mathbf{y}$.

Data collection plays a key role when estimating parameters for an epidemic model. For example, case data may be collected on a daily, weekly, or monthly basis. To determine how various time series data impact the practical identifiability of the model parameters, we explored four scenarios with different peaks of the epidemic curve and lengths of data collection. See Supplemental Figure 3 for plots of each scenario.

Scenario 1: The peak for the epidemic curve occurred at day 109 with the time span of 365 days, where $\beta = 0.0001$, $\gamma = 0.2$, and $\alpha = 0.03$.

Scenario 2: The peak for the epidemic curve occurred at day 25 with the time span of 50 days, where $\beta = 0.001$, $\gamma = 0.2$, and $\alpha = 0.03$.

Scenario 3: The peak for the epidemic curve occurred at day 109 with the time span of 100 days.

Scenario 4: The peak for the epidemic curve occurred at day 25 with the time span of 20 days.

Note that the time spans for Scenarios 1 and 2 were chosen to include roughly twice the time it takes for incidence (new infections) to drop below one. We selected this cutoff as a balance between capturing the entire prevalence curve and avoiding many observations near the steady state for incidence and cumulative incidence. Scenarios 3 and 4 simulate the effect of only having access to data before the peak of the epidemic.

3.1. Monte Carlo simulations

The history of Monte Carlo (MC) simulations can be traced back to the work of [57]. The MC method, as a sampling technique using random numbers and probability distributions, can be used to determine the practical identifiability of a model. This approach is due to its versatility and straightforward implementation. We perform MC simulations by generating M =10,000 synthetic data sets using the true parameter vector $\hat{\mathbf{p}}$ and adding noise to the data in increasing amounts. MC simulations are outlined in the following steps:

(1) Solve the SEIR ODEs (Eq 2.1) numerically with the true parameter vector $\hat{\mathbf{p}}$ to 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 =10,000 data sets with a given measurement error. We assume the error follows a normal distribution with mean 0 and variance $\sigma^2(t)$; that is, the data are described by

$$\mathbf{y}_{i,j} = g(\mathbf{x}(t_i), \hat{\mathbf{p}})(1 + \epsilon_{i,j}),$$

where $\epsilon \sim \mathcal{N}(0, \sigma)$ at the discrete data time points $\{t_i\}_{i=1}^n$ for all $j = 1, 2, ..., M$ data sets.

(3) Estimate the parameter vector $\mathbf{p}_j$ by fitting the dynamical model to each of the M simulated data sets. This is achieved by minimizing the difference between model output and the data generated for the specific scenario:

$$\mathbf{p}_j \approx \min_{\mathbf{p}} \sum_{i=1}^{n} \frac{(\mathbf{y}_{i,j} - g(\mathbf{x}(t_i), \mathbf{p}))^2}{(g(\mathbf{x}(t_i), \mathbf{p}))^2}.$$

This optimization problem is solved in MATLAB R2021a using the built-in function `fminsearchbnd`, which is part of the Optimization toolbox. Since `fminsearchbnd` is a local solver, the optimized minimum value can be influenced by the starting point. To avoid issues related to the starting value, we use the true parameter values as the initial parameter starting point provided to `fminsearchbnd`. Furthermore, we used the optimization function `fmincon` to test the consistency of the results in the methodology. Both functions produced the same qualitative results with similar AREs (the detailed ARE results are provided in the Supplemental document). For the remainder of the manuscript, the results shown refer to AREs from `fminsearchbnd`.

(4) Calculate the average relative estimation error (ARE) for each parameter in the set $\mathbf{p}$ following [16]:

$$\text{ARE}(p^{(k)}) = 100\% \times \frac{1}{M} \sum_{j=1}^{M} \frac{\left|\hat{p}^{(k)} - p_j^{(k)}\right|}{\left|\hat{p}^{(k)}\right|}, \tag{3.2}$$

where $p^{(k)}$ is the k-th parameter in the set $\mathbf{p}$, $\hat{p}^{(k)}$ is the k-th parameter in the true parameter vector $\hat{\mathbf{p}}$, and $p_j^{(k)}$ is the k-th element of $\mathbf{p}_j$.

(5) Repeat steps 1 through 4, increasing the level of noise ($\sigma = 0, 1, 5, 10, 20, 30\%$).

Following the convention in [17], we assume that a given parameter is practically identifiable if the ARE of the parameter is less than or equal to the measurement error, σ.

3.2. Correlation matrix method

Although the MC simulation approach is easy to understand and simple to implement, the associated computational cost is high due to a large number of repetitions and the associated optimization costs. One alternative is to utilize the sensitivity matrix of the model to compute the coefficient matrix for parameters of interest [20, 58, 59]. This requires much less computation and is relatively simple if measurement errors follow an identical and independent distribution. The correlation matrix (CM) method assesses the correlation between estimated parameters using a matrix of output sensitivities to model parameters. If estimated parameters are highly correlated, then they are considered practically unidentifiable.

CM assessments of identifiability are local to a particular parameter. The assessment also depends on the type of model observation, frequency of data collection, and assumed distribution of measurement noise. However, CM does not incorporate realizations of noisy observations, as in the MC approach. Instead, these considerations are incorporated in the following steps:

(1) Solve the SEIR model sensitivities numerically, for model observations $\mathbf{g}(\mathbf{x}(t), \mathbf{p})$ at the discrete collection times $\{t_i\}_{i=1}^n$ with respect to each parameter in $\mathbf{p}$.

(a) For prevalence and cumulative incidence, this is obtained from the SEIR model sensitivity equations and the extracted solutions for x_3 (I) and x_5 (C):

$$\frac{d}{dt}\frac{\partial \mathbf{x}}{\partial \mathbf{p}} = \frac{\partial f}{\partial \mathbf{x}}\frac{\partial \mathbf{x}}{\partial \mathbf{p}} + \frac{\partial f}{\partial \mathbf{p}}.$$

(b) For incidence, this is obtained from numerically integrating the quantity

$$\frac{\partial}{\partial \mathbf{p}}\int_{t_i}^{t_{i+1}} \beta S I dt = \int_{t_i}^{t_{i+1}} \left[\beta S \frac{\partial I}{\partial \mathbf{p}} + \left(\beta \frac{\partial S}{\partial \mathbf{p}} + \frac{\partial \beta}{\partial \mathbf{p}} S\right) I\right] dt,$$

where $\frac{\partial I}{\partial \mathbf{p}}$ and $\frac{\partial S}{\partial \mathbf{p}}$ come from solutions to the sensitivity equations.

(2) Construct a sensitivity matrix where the i, j^{th} component corresponds to the sensitivity of the model output $\mathbf{g}(\mathbf{x}(t), \mathbf{p})$ at time t_i to the j^{th} parameter in $\mathbf{p}$. This is denoted by

$$F_{i,j} = \frac{\partial \mathbf{g}}{\partial p_j}(\mathbf{x}(t_i), \mathbf{p}).$$

(3) Compute the inverse of the weighted Fisher information matrix, $IM = (F^T W F)^{-1}$, where W is the diagonal weighted matrix for the least squares error. For the assumed distribution of error in Eq (3.1), the weights are $g(\mathbf{x}(t_i), \mathbf{p})$.

(4) Compute the correlation coefficients $\chi_{ij} = IM_{ij}/\sqrt{IM_{ii} IM_{jj}}$ for all $i, j = 1, 2, 3$ and $i \neq j$, corresponding to correlations between the three parameters in $\mathbf{p}$.

(5) If the correlation coefficients are below 0.9 for all parameter pairs, then the parameter is practically identifiable for the assumed model observations.

The choice of 0.9 for the identifiability threshold in step five is chosen to reflect that a value close to unity represents a non-identifiable parameter pair (that is, these parameters are perfectly correlated) [17]. However, one of the challenges of the CM approach is that the determination of identifiability is qualitative in nature, and to make it quantitative requires the user to choose a value for this threshold. We conduct the CM approach for all data types described at the start of Section 3, to assess the practical identifiability of the model for the true parameters. We also repeat this process for estimated parameters $\mathbf{p}_j$ obtained in Step 3 of the MC approach (Section 3.1). This allows us to assess the perceived identifiability of parameter estimates obtained from noisy data.

4. Results

In this section, we will examine the output of the MC and CM methods described in Section 3 when applied to the SEIR model described in Section 2. For all of the estimated parameters, we chose the search bounds in the `fminsearchbnd` algorithm to be $[0, 1]$. Although, in practice, we may not know the true bounds of the parameters, this interval encompasses biologically feasible values with respect to their epidemiological definitions. For this model, the upper search bound corresponds to a highest possible transmission rate of 1000 new infections per infected individual each day and the shortest possible exposure and recovery times of 1 day each. We are interested in how different factors impact practical identifiability results and whether these results are aligned for both methods (MC and CM). We do not expect to have a one-to-one relationship in each scenario; however, it is beneficial to examine when and why we obtain matching outcomes when they occur in the simulations.

4.1. Monte Carlo (MC) results

Using the MC algorithm in Section 3.1, we tested the practical identifiability of the SEIR model parameters with respect to different data types, sampling frequencies, and time periods of available data. A model parameter is said to be *practically identifiable* if the ARE values given by Eq (3.2) are less than or equal to the noise percentage σ_0 for the given noise level. For example, if the ARE values for all levels of noise for β are less than the corresponding noise level, we say β is practically identifiable. If the ARE for β is more than its corresponding noise level for any noise level, we would say this parameter is not practically identifiable. If all parameters are practically identifiable in a given scenario, then the model is practically identifiable. This is consistent with the definition of practical identifiability used in [17]. The identifiability results from the MC approach are summarized in Table 2, and the details of the ARE values can be found in the Supplemental Tables 5–16.

Table 2. Practical identifiability results for all scenarios: prevalence, incidence, and cumulative incidence were using *fminsearchbnd* and *fmincon* with 10,000 iterations. Both methods yield the same identifiability results. A blue-shaded cell indicates that the parameter highlighted is practically identifiable.

Prevalence	Daily			Weekly			Monthly		
Scenario 1	🟦 β	🟦 γ	🟦 α	🟦 β	γ	🟦 α	🟦 β	γ	🟦 α
Scenario 2	β	γ	🟦 α	β	γ	🟦 α			
Scenario 3	🟦 β	γ	🟦 α	β	γ	α	β	γ	α
Scenario 4	β	γ	α						
Incidence	Daily			Weekly			Monthly		
Scenario 1	🟦 β	🟦 γ	🟦 α	🟦 β	γ	🟦 α	🟦 β	γ	🟦 α
Scenario 2	🟦 β	🟦 γ	🟦 α	🟦 β	🟦 γ	🟦 α			
Scenario 3	🟦 β	γ	🟦 α	🟦 β	γ	α	β	γ	α
Scenario 4	🟦 β	γ	α						
Cumulative Incidence	Daily			Weekly			Monthly		
Scenario 1	🟦 β	γ	α	β	γ	α	β	γ	α
Scenario 2	🟦 β	γ	α	β	γ	α			
Scenario 3	🟦 β	γ	α	🟦 β	γ	α	β	γ	α
Scenario 4	🟦 β	γ	α						

Before examining the ARE values, we generate violin plots for each scenario to observe the distribution of parameter values as noise is introduced in the MC procedure. Although we are not using the violin plots as a condition for practical identifiability, it does provide insight into how the values for each parameter change as we incorporate noise. Figure 2 represents the violin plots for prevalence, incidence, and cumulative incidence in Scenario 1 with respect to β. Predictably, as more noise is added to the data set, we observe a larger dispersion of parameter values for β, especially for 20% and 30% noise levels. The remaining plots for the other scenarios, data type, and parameters can be found in the Supplemental Figures 4, 5, 6, and 11.

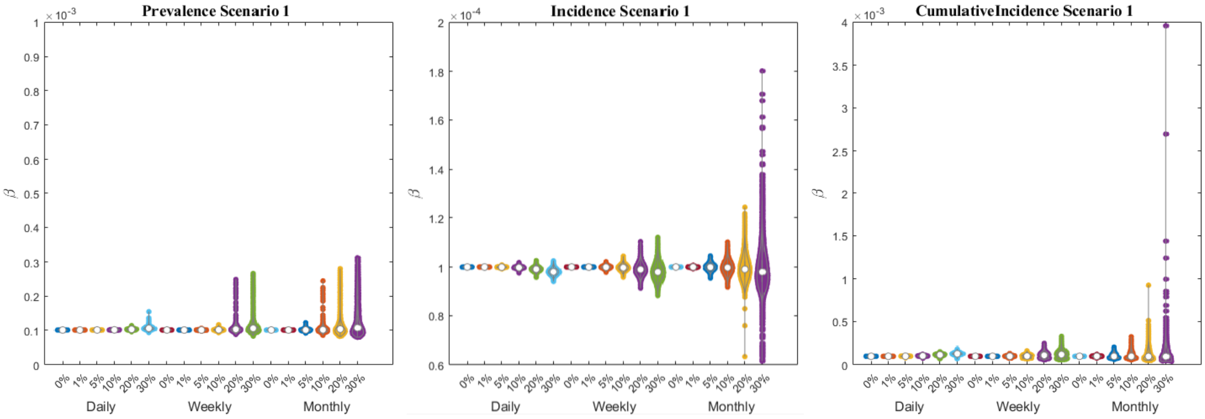


Figure 2. Violin Plots for prevalence, incidence, and cumulative incidence in Scenario 1.

As summarized in Table 2, a higher data density and longer temporal data availability tend to result in greater parameter identifiability overall. Moreover, the parameter γ had the fewest identifiable individual scenarios, with the highest number of identifiable scenarios being β and α depending on the disease data metric used. More scenarios of α were identifiable when using prevalence, while incidence resulted in more scenarios of β being identifiable. Additionally, the incidence scenarios had the most cases where parameters were identifiable, while only β is identifiable in some cases with cumulative incidence data. In particular, one special case where shorter temporal data yields better identifiability results occurred for parameter β with cumulative incidence data. This may be attributed to the performance of the optimization algorithm. It is also worth noting that comparing the weekly (monthly) results from Scenarios 1 and 2 with the daily (weekly) results from Scenarios 3 and 4, one could conclude that knowing more about the whole epidemic curve is more important for obtaining an identifiable model than obtaining dense data for a shorter period of time. A detailed discussion and reasoning about these findings are presented in Section 5.

4.2. *Correlation matrix (CM) results*

We next test the practical identifiability of the same scenarios as in the MC results, using the CM approach. Since CM reports identifiability of parameter pairs, instead of individual parameters, we holistically define a problem to be identifiable if all three parameter pairs have a correlation below 0.9. If any pair is more correlated, then the remaining estimates may be affected and result in an unidentifiable problem. In Supplemental Tables 29–31, we report these correlation results and, in Table 3, we summarize the resulting assessments of practical identifiability. We find that the most identifiable data type is incidence data, which leads to identifiable problems for any sampling rate in Scenario 1 and daily sampling rates in Scenarios 3 and 4. Estimating from prevalence data is an identifiable problem for daily and weekly sampling rates in Scenario 3, and problems using cumulative incidence are never identifiable.

Table 3. Practical identifiability for the CM approach for different scenarios and data types. A blue-shaded cell indicates that all parameters are identifiable for the given scenario and data type. Cumulative incidence is not shown, since all results were "not identifiable".

	Prevalence			Incidence		
	Daily	Weekly	Monthly	Daily	Weekly	Monthly
Scenario 1	No	No	No	Yes	Yes	Yes
Scenario 2	No	No		No	No	
Scenario 3	Yes	Yes	No	Yes	No	No
Scenario 4	No			Yes		

In most cases, parameter correlation values do not significantly change with sampling frequency. Only six correlation values increase by more than 0.1 between daily and monthly data, and none of these changes cross the 0.9 threshold for identifiability. Four correlation values increase past the 0.9 threshold between daily and monthly data, changing the assessment of identifiability for two data types (monthly prevalence and incidence data in Scenario 3). We therefore find two cases (out of nine) where increased sampling frequency improves the CM assessment of identifiability. However, this is a limited number of occurrences and it does not suggest an overall relationship between sampling frequency and identifiability under the CM criteria. In all other cases, the CM results are not sensitive to the frequency of data collection. Contrary to expectations, four correlation values decrease between daily and monthly data, although the change is less than 0.1 and does not cross the 0.9 threshold for identifiability.

4.3. Comparing MC and CM

In Supplemental Tables 32–34, we assess identifiability using the CM criteria for all estimates obtained in the MC process, under 0% and 30% noise levels, and then report the percentage of MC estimates that would be considered identifiable. At a 0% noise level, the CM results for the MC estimates match the results for the true parameters, since the estimates are very close to the true parameters. However, at a 30% noise level, the MC estimates may be far from the true parameter. For example, using daily prevalence data, the true parameter vector for Scenario 3 is identifiable by the CM approach. However, only 18% of the MC parameter estimates at 30% noise are identifiable by the CM criteria. Although β and α are identifiable by the MC criteria and remain close to their true values, there is a higher error in the estimate for γ. Because the CM results are affected by these incorrect γ values, the entire problem is considered unidentifiable. In other cases, incorrect MC estimates may be considered identifiable by the CM criteria. For example, using daily or weekly cumulative incidence data, the true parameter vector for Scenario 1 is not identifiable by the CM criteria. Under the MC criteria, β is identifiable using daily data, but is very close to the cutoff, and all other parameters are not identifiable. Using weekly data, none of the parameters are identifiable. However, 66% of the MC parameter estimates using daily data and 34% of the MC estimates using weekly data are identifiable by the CM criteria.

We next consider parameter correlations in four different cases, which correspond to all possible assessments of identifiability compared across the MC and CM criteria. The cases include daily prevalence data from Scenario 1, weekly prevalence data from Scenario 3, daily incidence data from Scenario 1, and daily cumulative incidence data from Scenario 1. In Supplemental Figure 13, we plot the normalized error of MC estimates using data with 30% noise, for estimated pairs of β:γ, β:α, and α:γ. We fit a straight line through the parameter error and report the slope, as a numerical assessment of correlation between MC parameter estimates. We consider two parameters correlated if the lines are sufficiently different from a strictly horizontal or vertical fit, which we define to be when the magnitude of the slope is between 0.5 and 2. This threshold is selected for symmetry around slopes with magnitude 1 and bounded away from horizontal and vertical lines, which would indicate no correlation. We select the specific bounds such that the two cases identifiable under the MC criteria also have uncorrelated parameter estimates. We compare these values to the CM results and present the correlations and identifiability in Table 36. We find that the sign of the correlation matches across the MC estimates and CM calculations in all but one case (α:γ for daily cumulative incidence data from Scenario 1). However, there is not a consistent match in the correlation magnitudes across MC estimates and CM calculations. The overall classification of "correlated" matches for nine (out of 12) pairs, but the thresholds for both criteria are somewhat arbitrary. For example, in Supplemental Tables 35 and 36, we assume wider and narrower ranges for MC slopes to be classified as "correlated", where the overall classification of "correlated" only matches in eight or seven cases.

Table 4. Parameter correlations for four cases, corresponding to different assessments of identifiability. For the MC approach, the "correlation" is the slope of a best-fit line between parameter estimates. For the CM approach, the correlation comes from calculations in Section 4.2. A blue-shaded cell indicates that the two parameters are classified as "not correlated".

Scenario	Identifiability		Correlations					
	MC	CM	MC			CM		
			β:γ	β:α	α:γ	β:γ	β:α	α:γ
S1 Prevalence daily	yes	no	-2.86	-0.22	6.12	-0.98	-0.66	0.54
S3 Prevalence weekly	no	yes	-1.17	0.15	-1.27	-0.87	0.89	-0.57
S1 Incidence daily	yes	yes	-9.77	-0.20	2.51	-0.84	-0.14	0.64
S1 Cumulative daily	no	no	-1.83	0.38	0.10	-0.97	0.92	-0.81

5. Discussion

In this study, we systematically investigated the impact of different data types (prevalence, incidence, and cumulative incidence) and sampling frequencies (daily, weekly, and monthly) on the practical identifiability of the SEIR model parameters, using two identifiability methods: the Monte Carlo (MC) method and the correlation matrix (CM) method. We found that incidence data, sampled at a higher frequency, resulted in the greatest degree of identifiability. While more data (as obtained at a higher sampling frequency) should intuitively lead to greater identifiability (although some exceptions are discussed in more detail below), it is less clear why incidence data should lead to greater identifiability compared to cumulative data. As such, we examine more closely the structure of these data types, and how they relate to model parameters. The disease metrics incidence and cumulative incidence are related to the flow into the exposed compartment, E, which depends on the transmission rate parameter β. On the other hand, prevalence is the result of integrating over the flows into and out of the infectious class, I, which are directly impacted by disease-progression rate γ and recovery rate α, and only indirectly by

β. Because cumulative incidence "smooths out" the information encoded in incidence, it is possible that different combinations of disease progression and recovery rates can yield very similar cumulative incidence curves. For example, if a decrease in the incubation period causes an increased rate of flow into the infected class, this could be balanced out by an increase in the recovery rate out of the infected class and thus lead to similar values of the cumulative incidence at the current time. This hypothesis is further supported by the transmission rate being the only identifiable parameter for cumulative incidence.

A more significant contributor, however, to the unidentifiability of cumulative incidence data is how cumulative incidence is generated. We argue that the method for adding measurement error in the case of cumulative data is patently unrealistic. The error model presented in our work (motivated by other studies on identifiability in outbreak models) assumes that the mean of the measurement error is proportional to the solution curve $C(t)$, resulting in very large measurement errors late in the epidemic. A more realistic way to generate the bootstrapped cumulative data would be to add measurement error to the incidence data and accumulate the incidence to obtain the corresponding cumulative incidence data, which results in much smaller measurement errors later in the epidemic compared with the first approach described here–a result of incidence tapering off after the peak of the epidemic. The challenge with the second approach is that the accumulated incidence data produces measurement errors that are not independent across time.

There are identifiable parameters in each of the various levels of data collection frequency, but the same cannot be said for each scenario. That is, given the low enough noise in our data, we are able to identify at least one parameter for each frequency of data collection, but we find that supplying data for only a portion of an epidemic is a limiting factor for practical identifiability. This is evident in Scenarios 3 and 4, where truncated time series data is used in the estimation process, resulting in few, if any, practically identifiable parameters. Since these scenarios mimic what would occur during an outbreak, care should be taken when estimating parameters before the epidemic has reached its peak, as there is likely to be a great deal of uncertainty in the estimates, and therefore, in the trajectories that follow. Projections of hypothetical control problems should take this uncertainty into consideration.

There are more identifiable problems by the CM criteria than problems where all parameters are identifiable by the MC criteria. However, CM does not allow for partially identifiable problems, and there are more cases where the MC approach classifies some parameters as independently identifiable than fully identifiable problems under the CM approach. The MC and CM assessments of "completely identifiable" problems seldom match, except for Scenario 1 daily incidence data. The assessments do, however, agree on the "completely unidentifiable" problems for Scenarios 1 and 2 using weekly or monthly cumulative incidence data, Scenario 3 monthly data of all types, and Scenario 4 daily prevalence data. For both assessments, we find that problems using incidence data are most often identifiable and problems using cumulative incidence data are least often identifiable. The assessments do not appear to follow similar patterns across sampling rates or cutoff dates. The MC approach more often follows the pattern that increased sampling rates (daily versus monthly) or longer time series (Scenario 3 versus Scenario 1) lead to more identifiable problems, while CM is less sensitive to sampling rates and is inconsistent in its response to longer time series.

Overall, we find that more data does correspond to more identifiable problems for the MC approach or does not affect identifiable problems for the CM approach. However, contrary to expectations, we find some cases where recovering parameters using shorter time series (Scenarios 3 or 4) are identifiable problems, where problems using longer time series (Scenarios 1 or 2) were not identifiable. Under the MC criteria, when using weekly cumulative incidence data, β is identifiable for Scenario 3 but not Scenario 1. This also occurs under the CM criteria using daily incidence data, where Scenario 4 is an identifiable problem but Scenario 2 is not. These results suggest that long tails without change in the time series may reduce identifiability. For the MC approach, the long tails may present an opportunity for the minimization to "fit to noise". This is supported by the difference across error levels, where the estimates have more error at lower noise levels for Scenario 3 (Supplemental Table 15) but as noise increases, the estimates for Scenario 1 have more error (Supplemental Table 13). Intuitively, this also makes sense for the CM approach, which relies on the model's sensitivity matrix; once the model solution reaches zero or levels off, there is less sensitivity to the model parameters. This explanation does not hold for prevalence data, which does not level off in our scenarios, but we find that under the CM criteria, when using daily or weekly prevalence data, Scenario 3 is an identifiable problem but Scenario 1 is not. We do not have an explanation for this outcome, except that Scenario 3 is borderline unidentifiable (0.89 correlation, in Supplemental Table 29) and that the 0.9 cutoff for identifiability may not be appropriate.

In Section 4.3, we further examine how MC and CM assessments of identifiability are related. In general, we found little overlap between the two, and by applying the CM criteria to MC estimates, we found that the CM criteria can lead to misleading results about parameter estimates. Most importantly, we found that parameter estimates that did not meet MC criteria for identifiability were sometimes identifiable by the CM criteria. This indicates that in practice, it is possible to estimate incorrect values for unidentifiable parameters but conclude by the CM criteria that the recovered parameters were identifiable. In other cases, the true parameter vector was identifiable by the CM criteria and the MC remained close to the true parameters, but the MC estimates were unidentifiable by the CM criteria. This discrepancy would also carry over to sensitivity-based confidence intervals, making it possible to have low confidence in a well-estimated parameter due to relatively small changes in the parameter space. However, it is important to note that our MC results represent a best case, in which numerical minimization starts from the true parameter estimate; in practice, the high sensitivities that affect the CM approach may affect estimates that start from other parts of the parameter space. In assessing the correlations of MC estimates, we found that the sign of correlation between parameters matched the results from the CM approach, but there was not a clear relationship between the magnitude of correlation for both approaches.

For the MC method, we tested two optimizers, `fmincon` and `fminsearchbnd`, from MATLAB for numerically finding the optimal parameters. `fmincon` is a function typically used for constrained optimization where the parameter space is limited through equality and inequality constraints and the objective function is at least second-order differentiable. It uses an interior point algorithm to find the optimal solution by default unless otherwise specified. On the other hand, `fminsearchbnd` requires upper and lower bounds for the parameter spaces, but cannot handle any other constraints on the parameters. It uses the Nelder-Mead simplex direct search algorithm, which does not require differentiability of the objective function. We implemented these two functions not to directly compare the numerical results, but to determine if they provided similar identifiability outputs.

We computed the AREs six times with different optimization algorithms, number of iterations, and data generation processes for each of the 27 cases (9 cases for each of the three data types as shown in Table 2) and presented the results for the run with the most iterations using `fminsearchbnd` and `fmincon` in the Supplemental Material. Notably, these computations exhibit both qualitative and quantitative variations. The first two times, we generate 1000 simulated data for each case, employing the `fmincon` and `fminsearchbnd` optimization methods separately. These two optimizers give different identifiability for 5 cases. To mitigate the stochastic variability arising from data generation, we computed the AREs two additional times using the same 1000 simulated data for each case across both optimizers. Then, the number of different cases is only one, though a different case than the previous 5. Furthermore, to reduce inter-case stochastic variability resulting from the utilization of partial data from Scenarios 1 and 2 in Scenarios 3 and 4, we conduct the final two rounds of computations. In these rounds, we initially generated 10,000 simulated data sets for Scenarios 1 and 2 and subsequently extracted the required data sets for all other cases. We observe that both optimizers yielded consistent qualitative outcomes across all cases. Based on this observation, we recommend increasing the number of simulated data sets whenever feasible, in order to enhance the reliability of the results.

Quantitatively, there are cases where AREs fluctuate around the 30% threshold over these six times. For example, the ARE corresponded to 30% noise data for β with prevalence as observation ranged from 28.3% to 30.9%. It would be difficult to assert practical identifiability in such cases. There are arguments that practical identifiability can be assessed as long as these thresholds are relatively on the same level. However, we found there is one case where the AREs ranged from 16% to 63%, while in another case, the AREs only vary from 16.1% to 17.1%. This poses a difficulty in the certainty of the results from the MC method. Our intensive computations suggest that the typically used 1000 simulations for the MC method may be insufficient in practice.

In addition, two distinct computational behaviors emerge when computing AREs for the same data: 1) In some cases, when the parameters are not identifiable, `fmincon` produces notably larger ARE than `fminsearchbnd;` 2) There are occurrences where the ARE computed from `fminsearchbnd` is nearly zero while the results from `fmincon` distinctly deviate from zero. The first phenomenon is likely due to the utilization of an unbounded search region by the `fmincon` optimizer, which arises from the absence of constraints on the parameter space. In contrast, when employing the `fminsearchbnd` optimizer, we impose a range of $[0, 1]$ for all parameters, thereby obtaining less AREs by confining the search within a bounded parameter space. The second phenomenon could be attributed to the search algorithm employed by these two optimizers. Our investigation indicates that, in instances where a non-zero ARE is obtained for zero noise data, the `fmincon` optimizer generates optimal parameter values that deviate from the accurate ones, even when the initial parameter values are set to the accurate values. Notably, the resulting parameters yield a larger error compared to the accurate values. This suggests that the algorithm encounters random deviations from the initial search point and is unable to converge back to the accurate parameters, potentially due to the intricate landscape of the error function.

6. Conclusions

In this study, we found that, for a range of parameter values, sampling rates, and criteria for practical identifiability, parameters estimated from incidence data are most often identifiable and parameters estimated from cumulative incidence data are least often identifiable. In practice, estimates obtained from cumulative incidence data should be treated with caution. However, it is important to note that identifiability was sensitive to underlying parameters across different scenarios. Additionally, we found that assessments of identifiability seldom agreed across the MC and CM approaches, with each method relying on cutoffs for identifiability that may be arbitrary. A further complication is that parameter estimation and identifiability results were sensitive to the choice of minimization algorithm, even after implementing common safeguards through numerical tolerances. Taken together, sensitivities to parameter values, the definition of identifiability, and numerical algorithms demonstrate that a single assessment of identifiability may be misleading or incomplete. We recommend a range of tests for identifiability to ensure confidence in the results. Finally, we note that our results are obtained for a relatively simple epidemic model under a best-case scenario, in which the model perfectly matches the processes generating the data, and were generated with knowledge of the underlying parameter vector, allowing for numerical searches to begin from the desired solutions. These conditions are impossible to meet in practice, and we still found fewer identifiable problems than expected. Parameter estimation for more complex models must be treated with significant caution and may require more rigorous testing of identifiability than is currently practiced. Although challenging, work in this area would greatly benefit from concrete guidelines for assessing practical identifiability from an array of methods.

Author contributions

Omar Saucedo, Amanda Laubmeier, Tingting Tang, Benjamin Levy, Lale Asik, Tim Pollington, Olivia Prosper Feldman: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Validation, Visualization, Roles/writing-original draft, Writing-review & editing. All authors have read and approved the final version of the manuscript for publication.

Use of AI tools declaration

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

Acknowledgments

This paper is based on work supported by the American Mathematical Society Mathematics Research Communities through National Science Foundation grants 1641020 and 1916439. OPF was supported in part by the National Science Foundation grant DMS-2045843.

Conflict of interest

The authors declare that they have no conflicts of interest.

References

1. M. Y. Li, H. L. Smith, L. Wang, Global dynamics of an SEIR epidemic model with vertical transmission, *SIAM J. Appl. Math.*, **62** (2001), 58–69. https://doi.org/10.1137/S0036139999359860
2. S. Gao, Y. Liu, J. J. Nieto, H. Andrade, Seasonality and mixed vaccination strategy in an epidemic model with vertical transmission, *Math. Comput. Simul.*, **81** (2011), 1855–1868. https://doi.org/10.1016/j.matcom.2010.10.032
3. A. d'Onofrio, Stability properties of pulse vaccination strategy in SEIR epidemic model, *Math. Biosci.*, **179** (2002), 57–72. https://doi.org/10.1016/S0025-5564(02)00095-0

4. P. Yan, S. Liu, Seir epidemic model with delay, *ANZIAM J.*, **48** (2006), 119–134. https://doi.org/10.1017/S144618110000345X

5. G. Li, Z. Jin, Global stability of a SEIR epidemic model with infectious force in latent, infected and immune period, *Chaos Solitons Fract.*, **25** (2005), 1177–1184. https://doi.org/10.1016/j.chaos.2004.11.062

6. J. Liu, Bifurcation analysis for a delayed SEIR epidemic model with saturated incidence and saturated treatment function, *J. Biol. Dynam.*, **13** (2019), 461–480. https://doi.org/10.1080/17513758.2019.1631965

7. R. Engbert, M. M. Rabe, R. Kliegl, S. Reich, Sequential data assimilation of the stochastic SEIR epidemic model for regional COVID-19 dynamics, *Bull. Math. Biol.*, **83** (2021), 1. https://doi.org/10.1007/s11538-020-00834-8

8. D. Efimov, R. Ushirobira, On an interval prediction of COVID-19 development based on a SEIR epidemic model, *Ann. Rev. Control*, **51** (2021), 477–487. https://doi.org/10.1016/j.arcontrol.2021.01.006

9. N. W. Ruktanonchai, J. Floyd, S. Lai, C. W. Ruktanonchai, A. Sadilek, P. Rente-Lourenco, et al., Assessing the impact of coordinated COVID-19 exit strategies across Europe, *Science*, **369** (2020), 1465–1470. https://doi.org/10.1126/science.abc5096

10. I. Borisov, E. Metelkin, Confidence intervals by constrained optimization-an algorithm and software package for practical identifiability analysis in systems biology, *PLOS Comput. Biol.*, **16** (2020), e1008495. https://doi.org/10.1371/journal.pcbi.1008495

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

12. C. Cobelli, J. J. DiStefano 3rd, Parameter and structural identifiability concepts and ambiguities: a critical review and analysis, *Amer. J. Physiol.-Reg. Integr. Compar. Physiol.*, **239** (1980), R7–R24. https://doi.org/10.1152/ajpregu.1980.239.1.R7

13. E. Walter, L. Pronzato, *Identification of Parametric Models: From Experimental Data*, Berlin: Springer Verlag, 1997.

14. A. Martynenko, A. Bück, *Intelligent Control in Drying*, New York: CRC Press, 2018.

15. F. G. Wieland, A. L. Hauber, M. Rosenblatt, C. Tönsing, J. Timmer, On structural and practical identifiability, *Current Opinion Syst. Biol*, **25** (2021), 60–69. https://doi.org/10.1016/j.coisb.2021.03.005

16. H. Miao, X. Xia, A. S. Perelson, H. Wu, On identifiability of nonlinear ODE models and applications in viral dynamics, *SIAM Rev.*, **53** (2011), 3–39. https://doi.org/10.1137/090757009

17. 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

18. M. Rodriguez-Fernandez, P. Mendes, J. R. Banga, A hybrid approach for efficient and robust parameter estimation in biochemical pathways, *Biosystems*, **83** (2006), 248–265. https://doi.org/10.1016/j.biosystems.2005.06.016

19. M. Rodriguez-Fernandez, J. A. Egea, J. R. Banga, Novel metaheuristic for parameter estimation in nonlinear dynamic biological systems, *BMC Bioinformatics*, **7** (2006), 483. https://doi.org/10.1186/1471-2105-7-483

20. H. T. Banks, S. Hu, W. C. Thompson, *Modeling and Inverse Problems in the Presence of Uncertainty*. New York: CRC Press, 2014.

21. D. Venzon, S. Moolgavkar, A method for computing profile-likelihood-based confidence intervals, *J. Royal Stat. Soc.: Ser. C (Applied Statistics)*, **37** (1988), 87–94. https://doi.org/10.2307/2347496

22. J. A. Jacquez, T. Perry, Parameter estimation: local identifiability of parameters, *Amer. J. Physiol.-Endocrinol. Metabol.*, **258** (1990), E727–E736.

23. A. Raue, C. Kreutz, T. Maiwald, J. Bachmann, M. Schilling, U. Klingmüller, et al., Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood, *Bioinformatics*, **25** (2009), 1923–1929. https://doi.org/10.1093/bioinformatics/btp358

24. M. C. Eisenberg, M. A. Hayashi, Determining identifiable parameter combinations using subset profiling, *Math. Biosci.*, **256** (2014), 116–126. https://doi.org/10.1016/j.mbs.2014.08.008

25. Y. H. Kao, M. C. Eisenberg, Practical unidentifiability of a simple vector-borne disease modeli: implications for parameter estimation and intervention assessment, *Epidemics*, **25** (2018), 89–100. https://doi.org/10.1016/j.epidem.2018.05.010

26. S. Baron, *Medical Microbiology*, 4 Eds., Galveston (TX): University of Texas Medical Branch at Galveston, 1996.

27. M. Lipsitch, T. Cohen, B. Cooper, J. M. Robins, S. Ma, L. James, et al., Transmission dynamics and control of severe acute respiratory syndrome, *Science*, **300** (2003), 1966–1970. https://doi.org/10.1126/science.1086616

28. S. Riley, C. A. Donnelly, N. M. Ferguson, Robust parameter estimation techniques for stochastic within-host macroparasite models, *J. Theoret. Biol.*, **225** (2003), 419–430. https://doi.org/10.1016/S0022-5193(03)00266-2

29. C. Fraser, C. A. Donnelly, S. Cauchemez, W. P. Hanage, M. D. Van Kerkhove, T. D. Hollingsworth, et al., Pandemic potential of a strain of influenza A (H1N1): early findings, *Science*, **324** (2009), 1557–1561. https://doi.org/10.1126/science.1176062

30. A. R. Tuite, J. Tien, M. Eisenberg, D. J. Earn, J. Ma, D. N. Fisman, Cholera epidemic in Haiti, 2010: using a transmission model to explain spatial spread of disease and identify optimal control interventions, *Ann. Int. Medic.*, **154** (2011), 593–601. https://doi.org/10.7326/0003-4819-154-9-201105030-00334

31. G. Chowell, H. Nishiura, Transmission dynamics and control of Ebola virus disease (EVD): a review, *BMC Med.*, **12** (2014), 196. https://doi.org/10.1186/s12916-014-0196-0

32. 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 Currents*, 2014. https://doi.org/10.1371/2Fcurrents.outbreaks.89c0d3783f36958d96ebbae97348d571

33. F. Brauer, P. Van den Driessche, J. Wu, L. J. Allen, *Mathematical Epidemiology*, Berlin: Springer, 2008.

34. L. Denis-Vidal, G. Joly-Blanchard, C. Noiret, Some effective approaches to check the identifiability of uncontrolled nonlinear systems, *Math. Comput. Simul.*, **57** (2001), 35–44. https://doi.org/10.1016/S0378-4754(01)00274-9

35. E. Walter, I. Braems, L. Jaulin, M. Kieffer, Guaranteed numerical computation as an alternative to computer algebra for testing models for identifiability, In: *Numerical Software with Result Verification. Lecture Notes in Computer Science*, Berlin: Springer, 2004.

36. S. Vajda, K. R. Godfrey, H. Rabitz, Similarity transformation approach to identifiability analysis of nonlinear compartmental models, *Math. Biosci.*, **93** (1989), 217–248. https://doi.org/10.1016/0025-5564(89)90024-2

37. N. D. Evans, M. J. Chapman, M. J. Chappell, K. R. Godfrey, Identifiability of uncontrolled nonlinear rational systems, *Automatica*, **38** (2002), 1799–1805. https://doi.org/10.1016/S0005-1098(02)00094-8

38. J. W. Yates, N. D. Evans, M. J. Chappell, Structural identifiability analysis via symmetries of differential equations, *Automatica*, **45** (2009), 2585–2591. https://doi.org/10.1016/j.automatica.2009.07.009

39. H. Pohjanpalo, System identifiability based on the power series expansion of the solution, *Math. Biosci.*, **41** (1978), 21–33. https://doi.org/10.1016/0025-5564(78)90063-9

40. G. Bellu, M. P. Saccomani, S. Audoly, L. D'Angiò, DAISY: a new software tool to test global identifiability of biological and physiological systems, *Comput. Meth. Programs Biomed*, **88** (2007), 52–61. https://doi.org/10.1016/j.cmpb.2007.07.002

41. L. Ljung, T. Glad, On global identifiability for arbitrary model parametrizations, *Automatica*, **30** (1994), 265–276. https://doi.org/10.1016/0005-1098(94)90029-9

42. E. Walter, Y. Lecourtier, Unidentifiable compartmental models: what to do? *Math. Biosci.*, **56** (1981), 1–25. https://doi.org/10.1016/0025-5564(81)90025-0

43. X. Xia, C. H. Moog, Identifiability of nonlinear systems with application to HIV/AIDS models, *IEEE Trans. Automat. Control*, **48** (2003), 330–336. https://doi.org/10.1109/TAC.2002.808494

44. A. F. Villaverde, A. Barreiro, A. Papachristodoulou, Structural identifiability of dynamic systems biology models, *PLoS Comput. Biol.*, **12** (2016), e1005153. https://doi.org/10.1371/journal.pcbi.1005153

45. A. Sedoglavic, A probabilistic algorithm to test local algebraic observability in polynomial time, *J. Symbol. Comput.*, **33** (2002), 735–755.

46. M. Anguelova, J. Karlsson,M. Jirstrand, Minimal output sets for identifiability, *Math. Biosci.*, **239** (2012), 139–153. https://doi.org/10.1016/j.mbs.2012.04.005

47. N. Meshkat, C. E. Kuo, J. DiStefano III, On finding and using identifiable parameter combinations in nonlinear dynamic systems biology models and COMBOS: a novel web implementation, *PLoS One*, **9** (2014), e110261. https://doi.org/10.1371/journal.pone.0110261

48. A. Raue, B. Steiert, M. Schelker, C. Kreutz, T. Maiwald, H. Hass, et al., Data2dynamics: a modeling environment tailored to parameter estimation in dynamical systems, *Bioinformatics*, **31** (2015), 3558–3560. https://doi.org/10.1093/bioinformatics/btv405

49. T. S. Ligon, F. Fröhlich, O. T. Chiş, J. R. Banga, E. Balsa-Canto, J. Hasenauer, GenSSI 2.0: multi-experiment structural identifiability analysis of SBML models, *Bioinformatics*, **34** (2018), 1421–1423. https://doi.org/10.1093/bioinformatics/btx735

50. H. Hong, A. Ovchinnikov, G. Pogudin, C. Yap, SIAN: software for structural identifiability analysis of ODE models, *Bioinformatics*, **35** (2019), 2873–2874. https://doi.org/10.1093/bioinformatics/bty1069

51. R. Dong, C. Goodbrake, H. A. Harrington, G. Pogudin, Differential elimination for dynamical models via projections with applications to structural identifiability, *SIAM J. Appl. Algebra Geometry*, **7** (2023), 194–235. https://doi.org/10.1137/22M1469067

52. X. Rey Barreiro, A. F. Villaverde, Benchmarking tools for a priori identifiability analysis, *Bioinformatics*, **39** (2023), btad065. https://doi.org/10.1093/bioinformatics/btad065

53. D. P. Lizarralde-Bejarano, D. Rojas-Díaz, S. Arboleda-Sánchez, M. E. Puerta-Yepes, Sensitivity, uncertainty and identifiability analyses to define a dengue transmission model with real data of an endemic municipality of Colombia, *PLoS One*, **15** (2020), e0229668. https://doi.org/10.1371/journal.pone.0229668

54. H. Wu, H. Zhu, H. Miao, A. S. Perelson, Parameter identifiability and estimation of hiv/aids dynamic models, *Bull. Math. Biol.*, **70** (2008), 785–799. https://doi.org/10.1007/s11538-007-9279-9

55. M. C. Eisenberg, S. L. Robertson, J. H. Tien, Identifiability and estimation of multiple transmission pathways in cholera and waterborne disease, *J. Theor. Biol.*, **324** (2013), 84–102. https://doi.org/10.1016/j.jtbi.2012.12.021

56. K. Roosa, G. Chowell, Assessing parameter identifiability in compartmental dynamic models using a computational approach: application to infectious disease transmission models, *Theor. Biol. Med. Model.*, **16** (2019), 1. https://doi.org/10.1186/s12976-018-0097-6

57. N. Metropolis, S. Ulam, The monte carlo method, *J. Amer. Stat. Assoc.*, **44** (1949), 335–341.

58. H. T. Banks, K. Holm, D. Robbins, Standard error computations for uncertainty quantification in inverse problems: Asymptotic theory vs. bootstrap, *Math. Comput. Model.*, **52** (2010), 1610–1625. https://doi.org/10.1016/j.mcm.2010.06.026

59. J. A. Jacquez, P. Greif, Numerical parameter identifiability and estimability: integrating identifiability, estimability, and optimal sampling design, *Math. Biosci.*, **77** (1985), 201–227. https://doi.org/10.1016/0025-5564(85)90098-7

Supplementary materials

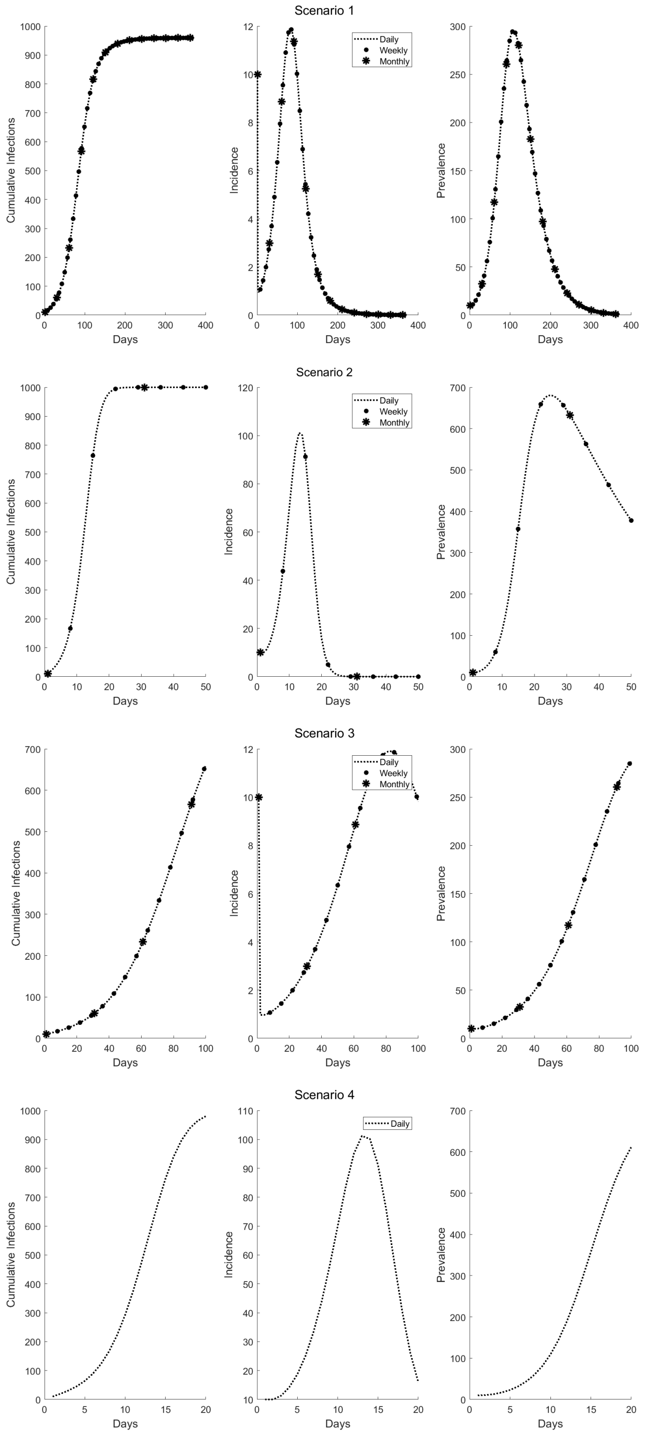


Figure 3. Plots for each scenario. In Scenario 1, the peak occurs on day 109 with the time span of 365 days. In Scenario 2, the peak occurs on day 25 with the time span of 50 days. In Scenario 3, the peak still occurs on day 109, but the time span is reduced to 100 days. Similarly, in Scenario 4, the peak is at day 25 with a time span of 20 days.

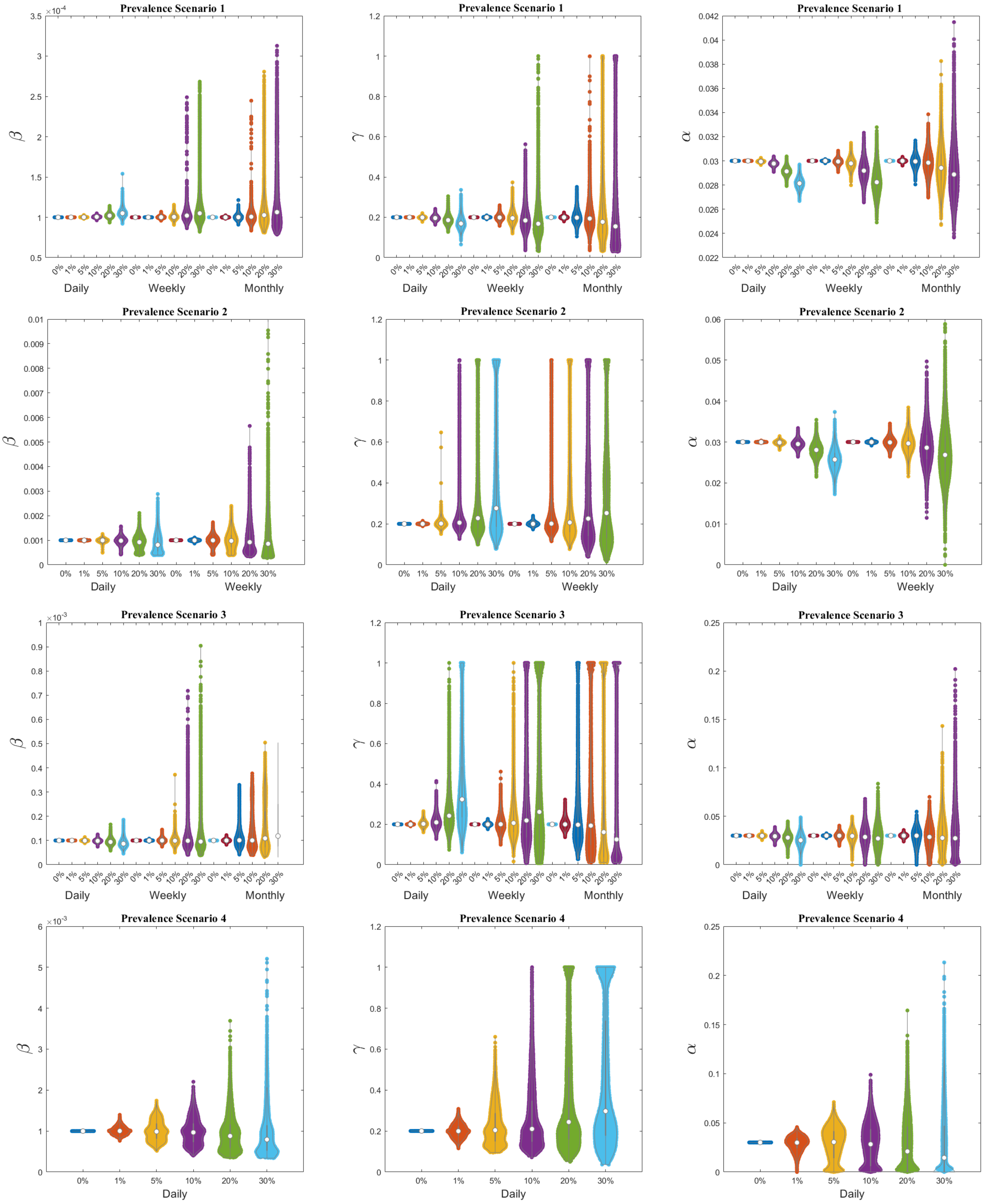


Figure 4. Violin plots for β, γ, and α using prevalence data for all four scenarios. The first column presents results for β, the second column for γ, and the third column for α. These distributions are generated from the MC algorithm using 10,000 iterations.

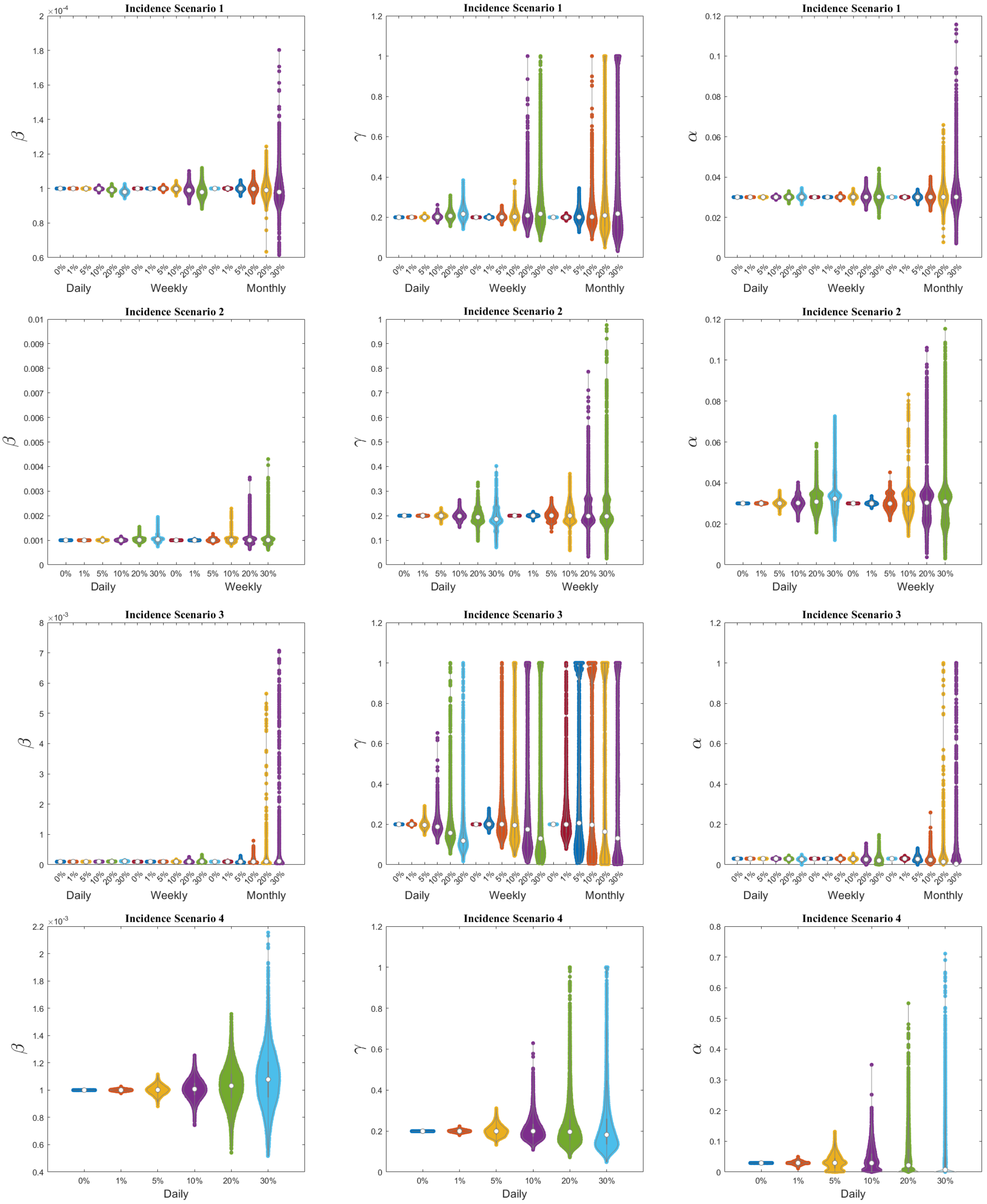


Figure 5. Violin plots for β, γ, and α using incidence data for all four scenarios. The first column presents results for β, the second column for γ, and the third column for α. These distributions are generated from the MC algorithm using 10,000 iterations.

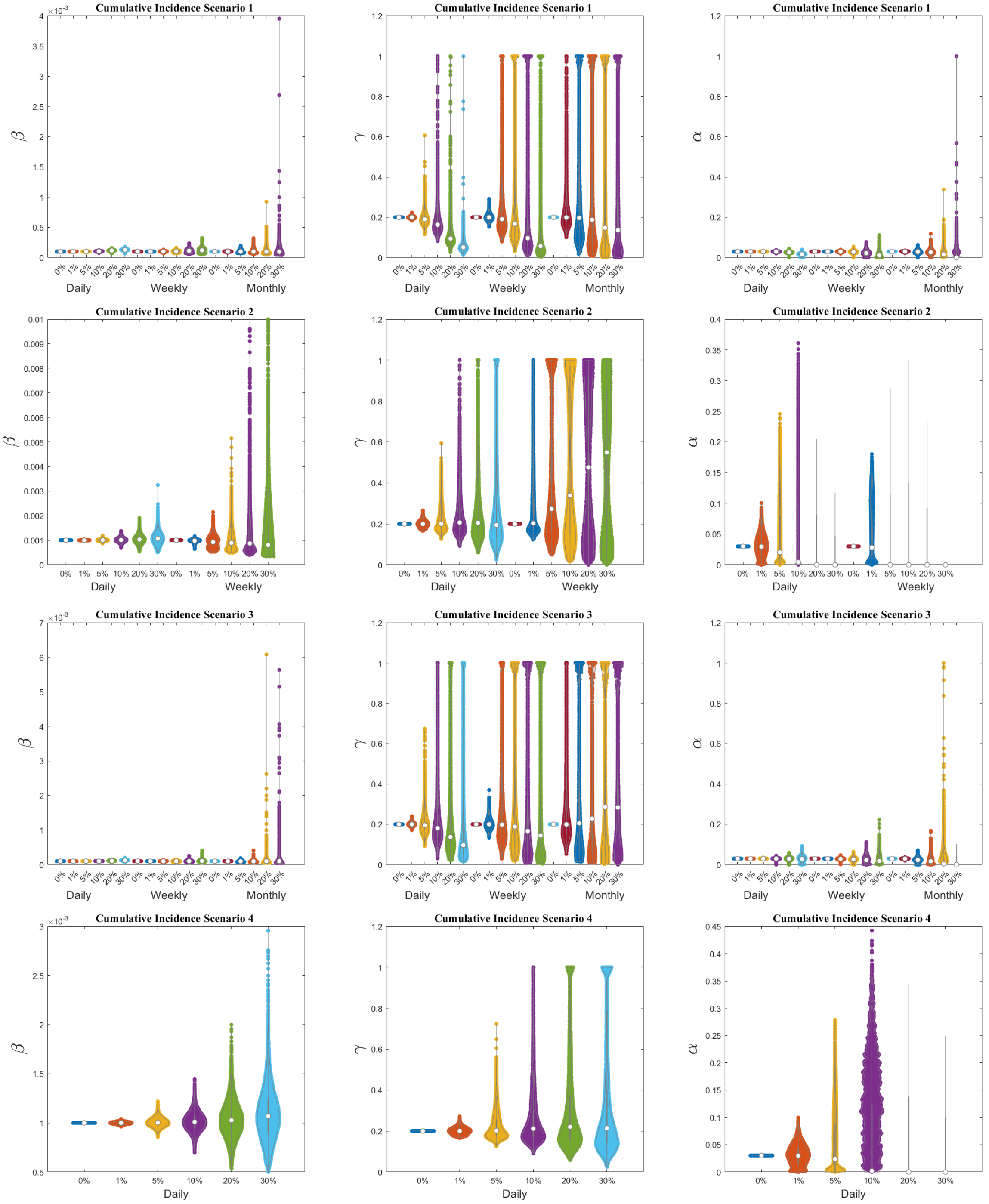


Figure 6. Violin plots for β, γ, and α using cumulative incidence data for all four scenarios. The first column presents results for β, the second column for γ, and the third column for α. These distributions are generated from the MC algorithm using 10,000 iterations.

Figures 7 through 10 plot normalized objective function value (fval) and normalized ARE for a given iteration versus other metrics of instance. Normalized ARE for a given scenario was obtained by dividing the given ARE value by the noise level for the scenario. Thus normalized ARE values below 1 are considered identifiable while values greater than 1 are unidentifiable. The normalized fval was calculated as

$$\frac{fval_{\text{True Params}} - fval_{\textbf{Est Params}}}{fval_{\text{True Params}}}.$$

With this in mind, positive values indicate the estimated parameter values fit the data better than the true parameter values while negative values indicate the true parameters provided a lower objective function value.

Figure 7 shows ARE values plotted on a log scale against our normalized objective function value (fval). The data is further stratified by parameter type and noise level. The normalized objective function values tend to be above 0 indicating that the estimated parameters fit the noisy data better than the true parameter values. Scenarios with the lowest ARE and fval values with the smallest variation tend to correspond to scenarios with a larger number of data points. This indicates that the optimization procedure struggles to identify the true parameters when noise is added to a small number of data points. Increasing ARE values tend to be associated with scenarios that have fewer data points and fvals that take on a larger range of values.

We notice similar dynamics in Figure 8, where the number of data points on a log scale is plotted against the normalized fval for different noise levels and data types. The normalized fval tending to be above 0 indicates that the estimated parameter values provide a better fit relative to true parameter values. As the number of data points decreases, the fval shows a wider range, again demonstrating how adding noise to a small number of data points complicates the parameter estimation process.

Figures 9 and 10 plot the number of data points in a given scenario on the log scale versus ARE as broken down by parameter and noise level. ARE values below 1 are considered identifiable while values greater than 1 are unidentifiable. Beta clearly produced the lowest ARE values across all noise levels and data types. We see that as long as there are enough data points, beta is the most likely to be identifiable.

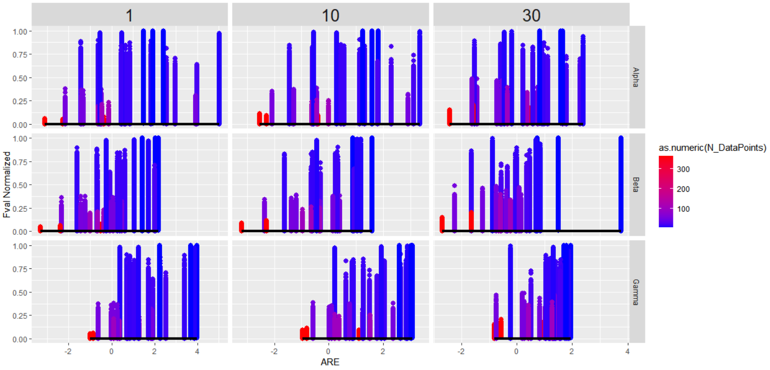


Figure 7. Normalized function values (fval) plotted against ARE on a log scale, for different noise levels across columns (1%, 10%, and 30%) and parameters across rows (α, β, and γ).

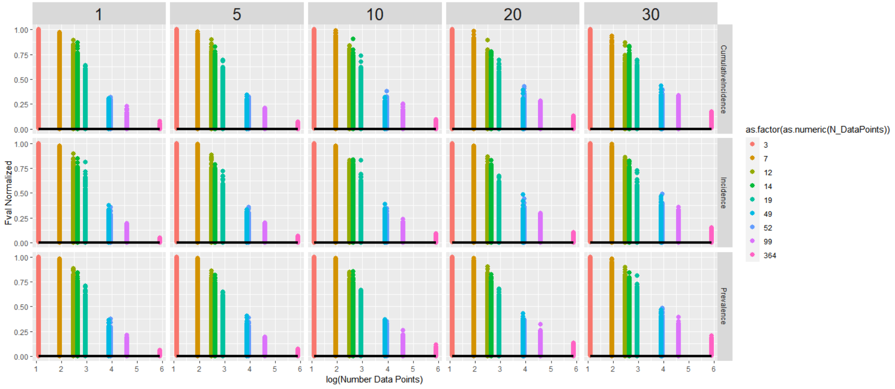


Figure 8. Normalized function values (fval) plotted against the number of data points on a log scale, for different noise levels across columns (1%, 5%, 10%, 20%, and 30%) and data types across rows (cumulative incidence, incidence, and prevalence)).

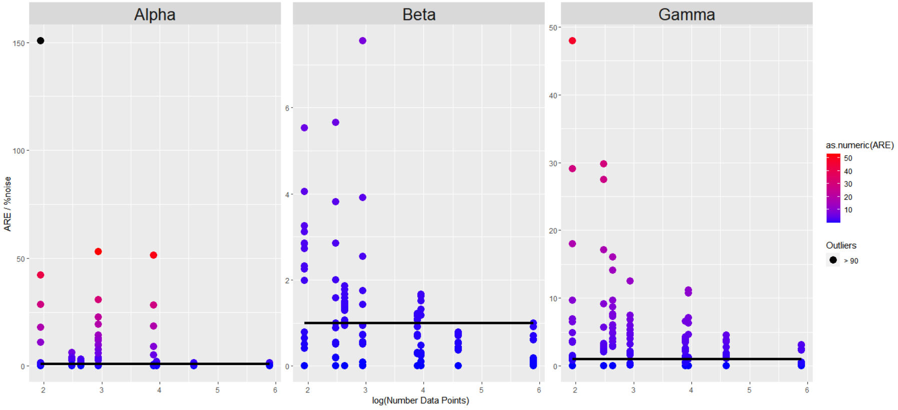


Figure 9. Normalized ARE values plotted against the number of data points on a log scale, for different parameters (α, β, and γ). Note that the scale of the vertical axis is different for each parameter.

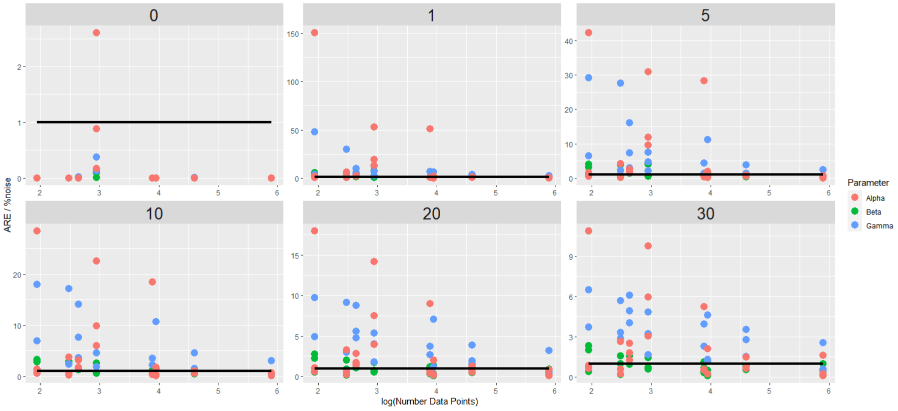


Figure 10. Normalized ARE values plotted against the number of data points on a log scale, for different noise levels (0%, 1%, 5%, 10%, 20%, and 30%). Note that the scale of the vertical axis is different for each noise level.

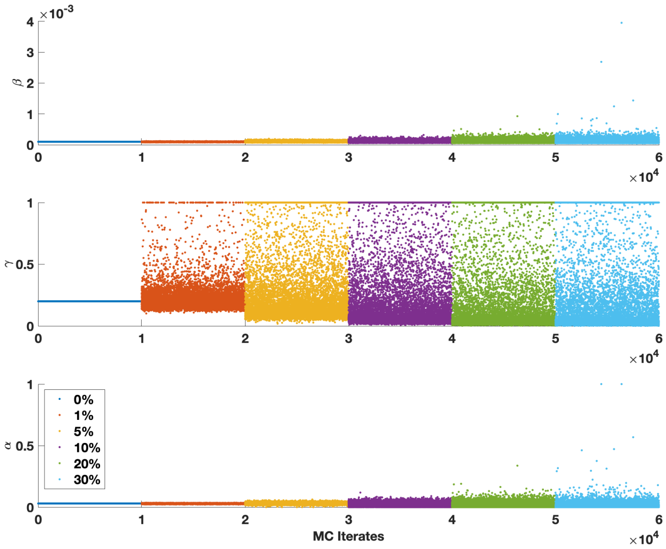


Figure 11. Estimates of β, γ, and α from 10,000 MC iterations at different noise levels (0%, 1%, 5%, 10%, 20%, and 30%) along the horizontal axis. Note that due to the relative magnitude, the scale of the vertical axis is different for β.

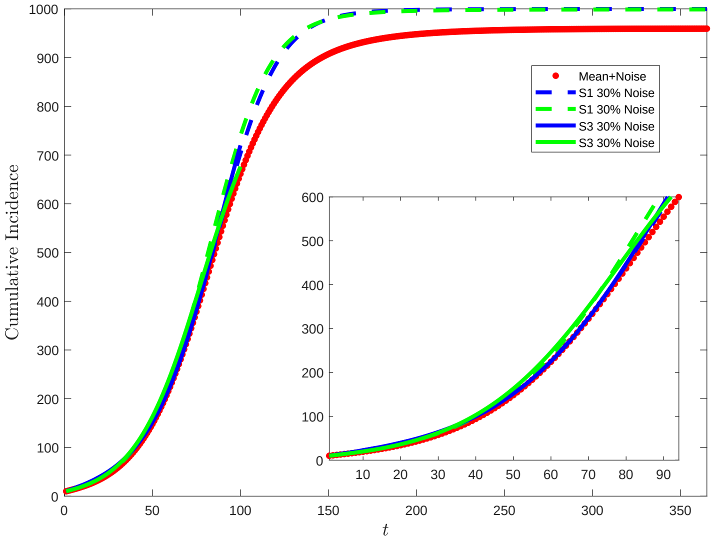
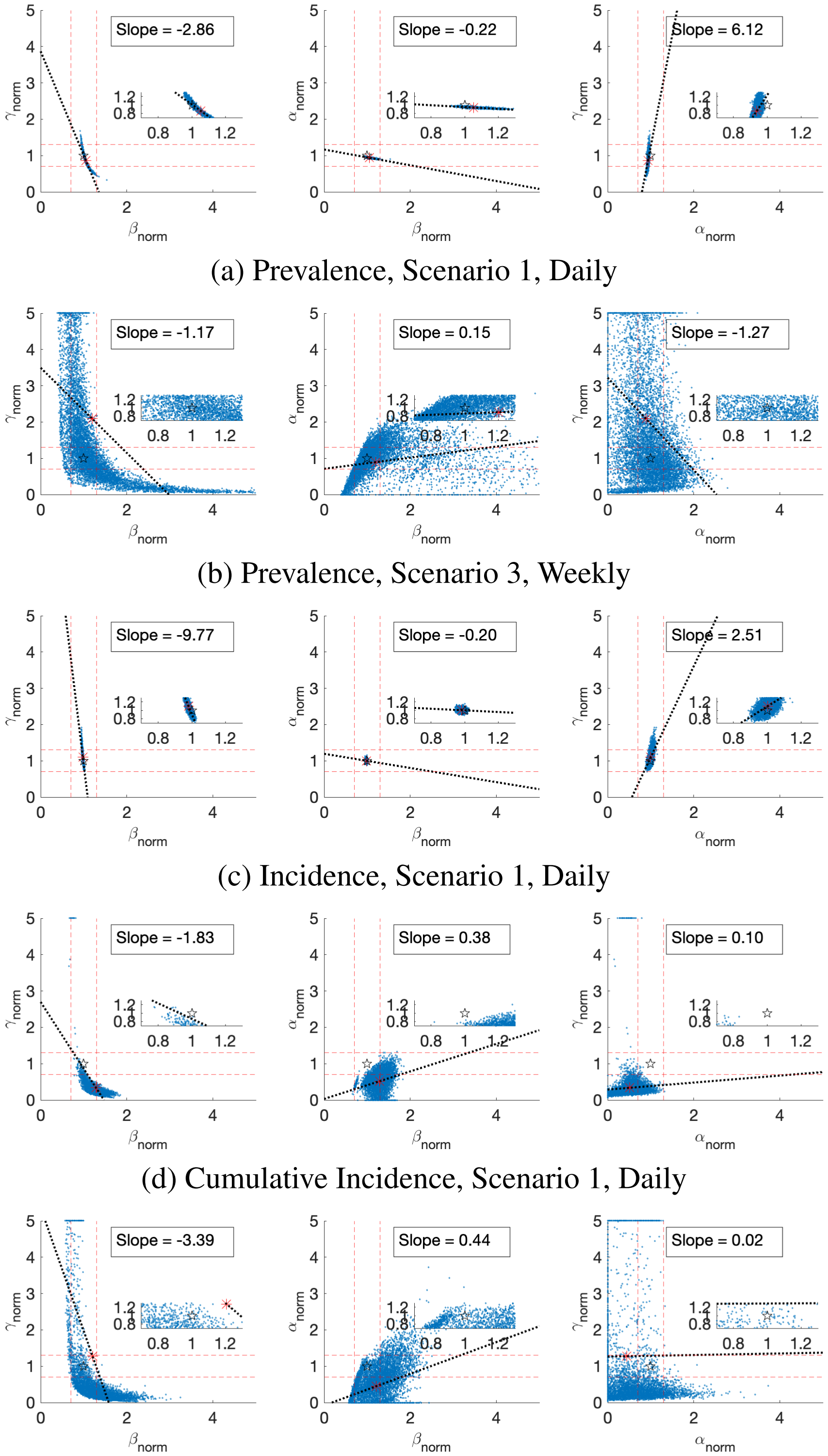


Figure 12. Comparison between the Scenario 1 and Scenario 3 model outputs when 30% noise is added to the data. This figure demonstrates that more information does not necessarily imply better fitting results.



(a) Prevalence, Scenario 1, Daily

(b) Prevalence, Scenario 3, Weekly

(c) Incidence, Scenario 1, Daily

(d) Cumulative Incidence, Scenario 1, Daily

(e) Cumulative Incidence, Scenario 1, Weekly

Figure 13. Scatter plot of pairs of normalized parameter estimates from MC simulations at a 30% noise level, for different scenarios and data types across rows. Note that parameter estimates are normalized by the true value for comparison across select scenarios. The linear fit through the scatter plot is indicated with a dashed black line.

Supplemental tables —Monte Carlo (MC) results from fminsearchbnd

Table 5. Average relative errors in parameter estimates from MC simulations using `fminsearchbnd` for *Prevalence data in Scenario 1*. The results are displayed for error levels ranging from 0% to 30% and various data sampling frequencies. The bottom row assesses identifiability, with blue-shaded cells indicating identifiable parameters.

Error level	Daily			Weekly			Monthly		
σ	β	γ	α	β	γ	α	β	γ	α
0 %	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1 %	0.09	0.42	0.04	0.24	1.10	0.11	0.51	2.37	0.24
5 %	0.47	2.16	0.26	1.21	5.55	0.59	2.50	11.64	1.18
10 %	1.03	4.45	0.79	2.50	11.20	1.25	5.52	24.69	2.36
20%	2.69	9.96	2.88	6.27	23.12	3.14	17.76	59.51	4.58
30 %	5.77	17.15	6.19	14.28	36.38	5.92	29.70	85.60	6.51
Identifiable	Yes	Yes	Yes	Yes	No	Yes	Yes	No	Yes

Table 6. Average relative errors in parameter estimates from MC simulations using `fminsearchbnd` for *Prevalence data in Scenario 2*. The results are displayed for error levels ranging from 0% to 30% and various data sampling frequencies. The bottom row shows the identifiability assessment, with blue-shaded cells indicating identifiable parameters.

Error level	Daily			Weekly		
σ	β	γ	α	β	γ	α
0 %	0.00	0.00	0.00	0.00	0.00	0.00
1 %	1.06	1.32	0.24	2.85	3.53	0.58
5 %	5.43	7.03	1.27	15.60	32.60	2.89
10%	11.69	21.42	2.76	28.37	69.06	5.80
20 %	24.40	74.11	7.25	45.12	98.03	12.26
30 %	32.74	117.37	14.49	59.74	110.75	19.18
Identifiable	No	No	Yes	No	No	Yes

Table 7. Average relative errors in parameter estimates from MC simulations using `fminsearchbnd` for *Prevalence data in Scenario 3*. The results are displayed for error levels ranging from 0% to 30% and various data sampling frequencies. The bottom row shows the identifiability assessment, with blue-shaded cells indicating identifiable parameters.

Error level	Daily			Weekly			Monthly		
σ	β	γ	α	β	γ	α	β	γ	α
0%	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1%	0.53	1.11	0.65	1.32	2.89	1.56	4.25	9.36	4.42
5%	2.63	5.75	3.24	6.69	15.01	7.86	23.00	70.17	21.24
10%	5.21	12.55	6.48	13.81	35.67	15.69	49.20	133.19	34.41
20%	10.72	38.33	13.45	33.24	95.75	28.93	76.36	168.39	52.20
30%	16.44	106.42	21.23	47.27	147.87	37.13	1287.14	178.02	68.08
Identifiable	Yes	No	Yes	No	No	No	No	No	No

Table 8. Average relative errors in parameter estimates from MC simulations using `fminsearchbnd` for *Prevalence data in Scenario 4*. The results are displayed for error levels ranging from 0% to 30% and various data sampling frequencies. The bottom row shows the identifiability assessment, with no parameters identified as identifiable.

Error level	Daily		
σ	β	γ	α
0%	0.09	0.14	0.17
1%	7.55	12.52	19.24
5%	19.57	37.17	48.05
10%	25.50	59.28	59.90
20%	34.91	106.81	78.40
30%	43.04	144.20	91.45
Identifiable	No	No	No

Table 9. Average relative errors in parameter estimates from MC simulations using `fminsearchbnd` for *Incidence data in Scenario 1*. The results are displayed for error levels ranging from 0% to 30% and various data sampling frequencies. The bottom row shows the identifiability assessment, with blue-shaded cells indicating identifiable parameters.

Error level	Daily			Weekly			Monthly		
σ	β	γ	α	β	γ	α	β	γ	α
0%	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1%	0.04	0.37	0.10	0.10	0.99	0.27	0.20	2.03	0.53
5%	0.19	1.87	0.52	0.48	4.92	1.36	0.98	10.38	2.70
10%	0.41	3.81	1.03	0.96	10.07	2.68	1.97	23.36	5.47
20%	1.03	8.07	1.97	1.99	22.32	5.13	3.75	61.51	10.80
30%	2.03	13.36	2.65	3.13	38.75	7.08	5.73	99.28	16.78
Identifiable	Yes	Yes	Yes	Yes	No	Yes	Yes	No	Yes

Table 10. Average relative errors in parameter estimates from MC simulations using `fminsearchbnd` for *Incidence data in Scenario 2*. The results are displayed for error levels ranging from 0% to 30% and various data sampling frequencies. The bottom row shows the identifiability assessment, with blue-shaded cells indicating identifiable parameters.

Error level	Daily			Weekly		
σ	β	γ	α	β	γ	α
0%	0.00	0.00	0.00	0.00	0.00	0.00
1%	0.29	0.54	0.61	0.78	1.49	1.60
5%	1.46	2.73	3.07	3.95	7.46	8.15
10%	3.07	5.64	6.45	6.45	12.44	13.25
20%	5.64	10.23	11.80	10.01	19.21	20.48
30%	2.03	13.36	2.65	12.36	24.00	24.94
Identifiable	Yes	Yes	Yes	Yes	Yes	Yes

Table 11. Average relative errors in parameter estimates from MC simulations using `fminsearchbnd` for *Incidence data in Scenario 3*. The results are displayed for error levels ranging from 0% to 30% and various data sampling frequencies. The bottom row shows the identifiability assessment, with blue-shaded cells indicating identifiable parameters.

Error level	Daily			Weekly			Monthly		
σ	β	γ	α	β	γ	α	β	γ	α
0%	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00
1%	0.37	1.45	0.54	1.50	5.65	1.89	8.91	40.28	11.26
5%	1.86	7.32	2.71	7.26	36.33	9.56	32.58	175.46	32.72
10 %	3.82	14.65	5.39	12.90	76.38	17.62	48.46	205.18	49.26
20%	8.73	30.68	11.01	21.30	111.74	33.87	75.24	193.70	90.65
30%	15.51	45.53	19.16	28.22	121.14	53.24	133.93	164.62	156.34
Identifiable	Yes	No	Yes	Yes	No	No	No	No	No

Table 12. Average relative errors in parameter estimates from MC simulations using `fminsearchbnd` for *Incidence data in Scenario 4*. The results are displayed for error levels ranging from 0% to 30% and various data sampling frequencies. The bottom row shows the identifiability assessment, with no parameters identified as identifiable.

Error level	Daily		
σ	β	γ	α
0%	2.26	3.93	5.46
1%	7.46	13.18	18.50
5%	19.16	38.24	45.39
10%	25.85	60.22	57.45
20%	36.36	103.55	76.31
30%	45.01	133.93	90.46
Identifiable	No	No	No

Table 13. Average relative errors in parameter estimates from MC simulations using `fminsearchbnd` for *Cumulative Incidence data in Scenario 1*. The results are displayed for error levels ranging from 0% to 30% and various data sampling frequencies. The bottom row shows the identifiability assessment, with blue-shaded cells indicating identifiable parameters.

Error level	Daily			Weekly			Monthly		
σ	β	γ	α	β	γ	α	β	γ	α
0%	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1%	0.60	2.34	0.69	1.62	6.34	1.83	5.65	29.82	6.21
5%	3.12	12.04	3.39	8.31	55.85	9.40	19.07	137.41	20.68
10 %	7.08	30.59	6.80	15.07	107.18	17.29	28.51	170.83	37.45
20 %	18.37	63.31	16.27	26.41	141.72	40.16	39.89	182.41	64.83
30%	29.67	76.41	48.44	35.10	138.42	63.45	47.30	170.29	79.50
Identifiable	Yes	No	No	No	No	No	No	No	No

Table 14. Average relative errors in parameter estimates from MC simulations using `fminsearchbnd` for *Cumulative Incidence data in Scenario 2*. The results are displayed for error levels ranging from 0% to 30% and various data sampling frequencies. The bottom row shows the identifiability assessment, with blue-shaded cells indicating identifiable parameters.

Error level	Daily			Weekly		
σ	β	γ	α	β	γ	α
0%	0.00	0.00	0.00	0.00	0.00	0.00
1%	0.92	6.59	51.18	5.53	47.91	150.65
5%	3.62	21.28	141.23	20.28	145.39	210.88
10%	7.19	34.65	183.97	32.61	179.75	284.48
20%	14.00	52.83	179.59	54.39	193.97	359.43
30%	20.50	68.05	156.92	69.78	195.01	326.25
Identifiable	Yes	No	No	No	No	No

Table 15. Average relative errors in parameter estimates from MC simulations using `fminsearchbnd` for *Cumulative Incidence data in Scenario 3*. The results are displayed for error levels ranging from 0% to 30% and various data sampling frequencies. The bottom row shows the identifiability assessment, with blue-shaded cells indicating identifiable parameters.

Error level	Daily			Weekly			Monthly		
σ	β	γ	α	β	γ	α	β	γ	α
0%	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1 %	0.70	3.62	0.89	1.86	9.72	2.33	7.60	54.17	6.95
5 %	3.50	19.11	4.53	8.90	80.27	13.36	19.24	188.96	37.96
10%	7.19	45.52	9.93	13.98	140.76	30.95	24.06	215.03	61.91
20 %	14.53	77.25	25.55	20.70	174.66	55.82	39.16	216.27	95.20
30 %	23.70	83.67	43.67	28.60	182.85	74.99	62.51	210.04	145.29
Identifiable	Yes	No	No	Yes	No	No	No	No	No

Table 16. Average relative errors in parameter estimates from MC simulations using `fminsearchbnd` for *Cumulative Incidence data in Scenario 4*. The results are displayed for error levels ranging from 0% to 30% and various data sampling frequencies. The bottom row shows the identifiability assessment, with blue-shaded cells indicating identifiable parameters.

Error level	Daily		
σ	β	γ	α
0%	0.01	0.12	0.88
1%	0.94	6.84	53.09
5%	3.67	23.59	154.27
10%	7.36	45.10	225.45
20 %	14.43	80.58	284.47
30 %	21.38	95.99	291.84
Identifiable	Yes	No	No

Supplemental tables —Monte Carlo (MC) results from `fmincon`

Table 17. Average relative errors in parameter estimates from MC simulations using `fmincon` for *Prevalence data in Scenario 1*. The results are displayed for error levels ranging from 0% to 30% and various data sampling frequencies. The bottom row assesses identifiability, with blue-shaded cells indicating identifiable parameters.

Error level	Daily			Weekly			Monthly		
σ	β	γ	α	β	γ	α	β	γ	α
0 %	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1 %	0.05	0.18	0.04	0.19	0.84	0.11	0.46	2.16	0.24
5 %	0.40	1.87	0.23	1.13	5.32	0.57	2.40	11.46	1.16
10 %	0.91	4.07	0.72	2.39	10.99	1.21	5.28	24.59	2.33
20%	2.47	9.38	2.79	5.99	22.86	3.07	17.08	59.54	4.55
30 %	5.39	16.33	6.10	13.51	36.05	5.85	28.85	85.52	6.49
Identifiable	Yes	Yes	Yes	Yes	No	Yes	Yes	No	Yes

Table 18. Average relative errors in parameter estimates from MC simulations using `fmincon` for *Prevalence data in Scenario 2*. The results are displayed for error levels ranging from 0% to 30% and various data sampling frequencies. The bottom row assesses identifiability, with blue-shaded cells indicating identifiable parameters.

Error level	Daily			Weekly		
σ	β	γ	α	β	γ	α
0 %	0.01	0.01	0.00	0.02	0.02	0.00
1 %	1.06	1.32	0.24	2.85	3.53	0.58
5 %	5.44	7.05	1.27	15.60	32.17	2.89
10%	11.70	21.41	2.76	28.35	67.98	5.80
20 %	24.41	73.83	7.25	45.01	95.99	12.26
30 %	32.74	116.67	14.49	69.30	106.77	19.20
Identifiable	No	No	Yes	No	No	Yes

Table 19. Average relative errors in parameter estimates from MC simulations using `fmincon` for *Prevalence data in Scenario 3*. The results are displayed for error levels ranging from 0% to 30% and various data sampling frequencies. The bottom row assesses identifiability, with blue-shaded cells indicating identifiable parameters.

Error level	Daily			Weekly			Monthly		
σ	β	γ	α	β	γ	α	β	γ	α
0%	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1%	0.50	0.91	0.66	1.36	3.01	1.62	3.85	8.93	4.24
5%	2.73	6.03	3.41	6.63	15.21	7.94	21.99	68.52	21.20
10%	5.38	13.17	6.76	13.70	36.26	15.88	47.07	129.59	34.81
20%	10.98	39.95	13.94	32.05	95.21	29.14	73.95	163.65	52.35
30%	16.80	108.70	21.93	45.91	145.96	37.28	2851.12	173.97	68.07
Identifiable	Yes	No	Yes	No	No	No	No	No	No

Table 20. Average relative errors in parameter estimates from MC simulations using `fmincon` for *Prevalence data in Scenario 4*. The results are displayed for error levels ranging from 0% to 30% and various data sampling frequencies. The bottom row assesses identifiability, with blue-shaded cells indicating identifiable parameters.

Error level	Daily		
σ	β	γ	α
0%	2.26	3.93	5.46
1%	7.46	13.18	18.50
5%	19.16	38.24	45.39
10%	25.85	60.22	57.45
20%	36.36	103.55	76.31
30%	45.01	133.93	90.46
Identifiable	No	No	No

Table 21. Average relative errors in parameter estimates from MC simulations using `fmincon` for *Incidence data in Scenario 1*. The results are displayed for error levels ranging from 0% to 30% and various data sampling frequencies. The bottom row assesses identifiability, with blue-shaded cells indicating identifiable parameters.

Error level	Daily			Weekly			Monthly		
σ	β	γ	α	β	γ	α	β	γ	α
0%	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1%	0.01	0.03	0.02	0.07	0.52	0.19	0.20	1.87	0.51
5%	0.20	1.65	0.48	0.49	4.90	1.35	0.97	10.49	2.71
10%	0.45	3.89	1.01	0.97	10.20	2.69	1.96	23.65	5.49
20%	1.10	8.42	1.97	2.01	22.71	5.13	3.72	61.78	10.80
30%	2.12	13.89	2.65	3.16	39.27	7.08	5.69	99.19	16.78
Identifiable	Yes	Yes	Yes	Yes	No	Yes	Yes	No	No

Table 22. Average relative errors in parameter estimates from MC simulations using `fmincon` for *Incidence data in Scenario 2*. The results are displayed for error levels ranging from 0% to 30% and various data sampling frequencies. The bottom row assesses identifiability, with blue-shaded cells indicating identifiable parameters.

Error level	Daily			Weekly		
σ	β	γ	α	β	γ	α
0%	0.00	0.00	0.00	0.00	0.00	0.00
1%	0.13	0.24	0.28	0.30	0.60	0.66
5%	0.56	1.08	1.26	1.70	3.31	3.71
10%	1.22	2.30	2.73	3.39	6.45	7.32
20%	2.78	5.01	6.11	6.55	11.57	13.58
30%	5.20	8.51	10.83	9.50	16.26	18.81
Identifiable	Yes	Yes	Yes	Yes	Yes	Yes

Table 23. Average relative errors in parameter estimates from MC simulations using `fmincon` for *Incidence data in Scenario 3*. The results are displayed for error levels ranging from 0% to 30% and various data sampling frequencies. The bottom row assesses identifiability, with blue-shaded cells indicating identifiable parameters.

Error level	Daily			Weekly			Monthly		
σ	β	γ	α	β	γ	α	β	γ	α
0%	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00
1%	0.33	1.33	0.55	1.44	5.63	1.92	5.61	30.30	7.85
5%	1.79	7.20	2.76	7.14	38.76	9.66	28.82	154.69	30.21
10 %	3.67	14.51	5.42	12.75	79.46	17.71	46.28	188.07	47.96
20%	8.47	30.64	11.08	21.02	111.28	33.81	68.55	182.10	85.63
30%	15.19	45.31	19.21	28.01	119.14	53.23	93.48	154.56	125.72
Identifiable	Yes	No	Yes	Yes	No	No	No	No	No

Table 24. Average relative errors in parameter estimates from MC simulations using `fmincon` for *Incidence data in Scenario 4*. The results are displayed for error levels ranging from 0% to 30% and various data sampling frequencies. The bottom row assesses identifiability, with blue-shaded cells indicating identifiable parameters.

Error level	Daily		
σ	β	γ	α
0%	0.01	0.00	0.00
1%	0.52	2.25	14.04
5%	2.57	10.02	58.51
10%	5.24	18.75	99.72
20%	10.85	36.77	155.21
30%	16.92	49.96	183.86
Identifiable	Yes	No	No

Table 25. Average relative errors in parameter estimates from MC simulations using `fmincon` for *Cumulative Incidence data in Scenario 1*. The results are displayed for error levels ranging from 0% to 30% and various data sampling frequencies. The bottom row assesses identifiability, with blue-shaded cells indicating identifiable parameters.

Error level	Daily			Weekly			Monthly		
σ	β	γ	α	β	γ	α	β	γ	α
0%	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1%	0.43	1.54	0.63	1.48	6.12	1.78	4.90	29.22	5.54
5%	2.69	10.82	3.36	7.90	54.30	9.24	18.30	131.12	20.40
10%	6.43	28.93	6.87	14.64	103.96	17.28	29.64	162.19	39.33
20%	17.75	61.80	16.55	26.46	138.19	40.81	47.06	170.81	72.01
30%	29.35	76.12	48.68	35.54	135.64	64.31	1839.19	157.51	98.54
Identifiable	Yes	No	No	No	No	No	No	No	No

Table 26. Average relative errors in parameter estimates from MC simulations using `fmincon` for *Cumulative Incidence data in Scenario 2*. The results are displayed for error levels ranging from 0% to 30% and various data sampling frequencies. The bottom row assesses identifiability, with blue-shaded cells indicating identifiable parameters.

Error level	Daily			Weekly		
σ	β	γ	α	β	γ	α
0%	1.23	13.89	105.71	0.12	1.89	9.89
1%	0.99	8.07	61.47	5.90	52.39	158.97
5%	3.66	21.79	139.17	20.27	143.20	226.75
10%	7.28	35.27	182.15	32.51	174.51	305.53
20%	14.12	53.18	179.58	173.07	187.73	389.63
30%	20.60	67.95	155.22	1977.03	188.83	413.24
Identifiable	Yes	No	No	No	No	No

Table 27. Average relative errors in parameter estimates from MC simulations using `fmincon` for *Cumulative Incidence data in Scenario 3*. The results are displayed for error levels ranging from 0% to 30% and various data sampling frequencies. The bottom row assesses identifiability, with blue-shaded cells indicating identifiable parameters.

Error level	Daily			Weekly			Monthly		
σ	β	γ	α	β	γ	α	β	γ	α
0%	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1 %	0.50	2.04	0.84	1.62	8.62	2.43	6.26	44.92	7.50
5 %	3.06	17.52	4.55	8.55	76.71	13.34	18.25	170.24	36.76
10%	6.65	43.63	9.82	13.67	135.67	30.67	23.80	201.17	60.30
20 %	13.96	75.08	25.29	20.52	169.93	55.42	38.55	206.35	92.92
30 %	23.18	82.04	43.28	28.47	178.60	74.76	1009.17	201.16	143.25
Identifiable	Yes	No	No	Yes	No	No	No	No	No

Table 28. Average relative errors in parameter estimates from MC simulations using `fmincon` for *Cumulative Incidence data in Scenario 4*. The results are displayed for error levels ranging from 0% to 30% and various data sampling frequencies. The bottom row assesses identifiability, with blue-shaded cells indicating identifiable parameters.

Error level	Daily		
σ	β	γ	α
0%	0.02	0.30	2.18
1%	1.01	8.24	62.80
5%	3.72	23.96	151.22
10%	7.44	45.34	221.31
20 %	14.54	80.70	284.35
30 %	21.53	96.19	296.40
Identifiable	Yes	No	No

Supplemental tables —correlation matrix results

The following tables contain the correlation coefficients for the different scenarios and data types, rounded to two decimal places. If the correlation coefficient is above 0.9 for any pair of parameters, the parameter vector is "not identifiable" for the scenario and data type.

Table 29. The correlation matrix identifiability results for *Prevalence data* are presented. Key correlation coefficients, those approaching or exceeding the 0.9 threshold for parameter identifiability, are highlighted in blue-shaded cells. For comparison, the identifiable parameters from the MCMC results are included in the final row.

Scenario	1			2		3			4
Sampling	D	W	M	D	W	D	W	M	D
β, γ correlation	-0.98	-0.98	-0.98	-0.99	-0.99	-0.86	-0.87	-0.95	-1.00
β, α correlation	-0.66	-0.67	-0.68	0.01	0.10	0.90	0.89	0.94	-0.97
γ, α correlation	0.54	0.54	0.56	-0.01	-0.10	-0.55	-0.57	-0.79	0.99
Identifiable	No	No	No	No	No	Yes	Yes	No	No
MCMC Results	β, γ, α	β, α	β, α	α	α	β, α	-	-	-

Table 30. The correlation matrix identifiability results for *Incidence data* are presented. Key correlation coefficients, those approaching or exceeding the 0.9 threshold for parameter identifiability, are highlighted in blue-shaded cells. For comparison, the identifiable parameters from the MCMC results are included in the final row.

Scenario	1			2		3			4
Sampling	D	W	M	D	W	D	W	M	D
β, γ correlation	-0.84	-0.84	-0.82	-1.00	-1.00	-0.90	-0.90	-0.97	-0.32
β, α correlation	-0.14	-0.14	-0.11	1.00	1.00	0.74	0.74	0.83	0.20
γ, α correlation	0.64	0.64	0.64	-0.99	-0.99	-0.38	-0.39	-0.69	0.86
Identifiable	Yes	Yes	Yes	No	No	Yes	No	No	Yes
MCMC Results	β, γ, α	β, α	β	β, γ, α	β, γ, α	β	-	-	-

Table 31. The correlation matrix identifiability results for *Cumulative Incidence data* are presented. Key correlation coefficients, those approaching or exceeding the 0.9 threshold for parameter identifiability, are highlighted in blue-shaded cells. For comparison, the identifiable parameters from the MCMC results are included in the final row.

Scenario	1			2		3			4
Sampling	D	W	M	D	W	D	W	M	D
β, γ correlation	-0.97	-0.97	-0.99	0.56	-0.68	-0.92	-0.92	-0.98	0.58
β, α correlation	0.92	0.92	0.97	0.73	-0.59	0.33	0.38	0.75	0.75
γ, α correlation	-0.81	-0.81	-0.93	0.97	0.99	0.06	0.00	-0.61	0.97
Identifiable	No	No	No	No	No	No	No	No	No
MCMC Results	-	-	-	-	-	-	-	-	-

The following tables report the percent of identifiable parameters from MCMC iterates. In three scenarios, we omit parameter estimates, which result in a noninvertible Fischer information matrix. This happens for monthly Prevalence data under Scenario 3 (984 parameters omitted), weekly Cumulative Incidence data under Scenario 2 (4 parameters omitted), and monthly Cumulative Incidence data under Scenario 2 (641 parameters omitted). The corresponding results are indicated with an asterisk in the tables below.

Table 32. Correlation matrix identifiability rates for MCMC parameter estimates, using *Prevalence data*. For comparison, we include results for true parameters in the last row. The asterisk indicates a case where parameter estimates were omitted due to noninvertible Fisher matrices.

Scenario	1			2		3			4
Sampling	D	W	M	D	W	D	W	M	D
0% Error Estimates	0%	0%	0%	0%	0%	100%	100%	0%	0%
30% Error Estimates	0%	0%	0%	0%	0.33%	18%	10%	5.2*%	1.5%
True Parameter Values	No	No	No	No	No	Yes	Yes	No	No

Table 33. Correlation matrix identifiability rates for MCMC parameter estimates, using *Incidence data*. For comparison, we include results for true parameters in the last row.

Scenario	1			2		3			4
Sampling	D	W	M	D	W	D	W	**M**	D
0% Error Estimates	100%	100%	100%	0%	0%	100%	0%	0%	100%
30% Error Estimates	100%	100%	96%	0%	0%	86%	45%	6.2%	63%
True Parameters	Yes	Yes	Yes	No	No	Yes	No	No	Yes

Table 34. Correlation matrix identifiability rates for MCMC parameter estimates, using *Cumulative Incidence data*. For comparison, we include results for true parameters in the last row. The asterisks indicate cases where parameter estimates were omitted due to noninvertible Fisher matrices.

Scenario	1			2		3			4
Sampling	D	W	M	D	W	D	W	M	D
0% Error Estimates	0%	0%	0%	100%	100%	0%	0%	0%	100%
30% Error Estimates	66%	34%	13%	17%	3*%	38%	15%	1.8*%	25%
True Parameters	No	No	No	No	No	No	No	No	No

Table 35. Parameter correlations for four cases, corresponding to different assessments of identifiability. For the MC approach, the "correlation" is the slope of a best-fit line between parameter estimates, where "correlated" requires a slope range between $0.2 < |m| < 5$. For the CM approach, the correlation comes from calculations in Section 4.2. A blue-shaded cell indicates that the two parameters are classified as "not correlated".

Scenario	Identifiability		Correlations					
	MC	CM	MC			CM		
			β:γ	β:α	α:γ	β:γ	β:α	α:γ
S1 Prevalence daily	yes	no	-2.86	-0.22	6.12	-0.98	-0.66	0.54
S3 Prevalence weekly	no	yes	-1.17	0.15	-1.27	-0.87	0.89	-0.57
S1 Incidence daily	yes	yes	-9.77	-0.20	2.51	-0.84	-0.14	0.64
S1 Cumulative daily	no	no	-1.83	0.38	0.10	-0.97	0.92	-0.81

Table 36. Parameter correlations for four cases, corresponding to different assessments of identifiability. For the MC approach, the "correlation" is the slope of a best-fit line between parameter estimates, where "correlated" requires a slope range between $0.8 < |m| < 1.25$. For the CM approach, the correlation comes from calculations in Section 4.2. A blue-shaded cell indicates that the two parameters are classified as "not correlated".

Scenario	Identifiability		Correlations					
	MC	CM	MC			CM		
			β:γ	β:α	α:γ	β:γ	β:α	α:γ
S1 Prevalence daily	yes	no	-2.86	-0.22	6.12	-0.98	-0.66	0.54
S3 Prevalence weekly	no	yes	-1.17	0.15	-1.27	-0.87	0.89	-0.57
S1 Incidence daily	yes	yes	-9.77	-0.20	2.51	-0.84	-0.14	0.64
S1 Cumulative daily	no	no	-1.83	0.38	0.10	-0.97	0.92	-0.81