ROYAL SOCIETY OPEN SCIENCE

royalsocietypublishing.org/journal/rsos

Research

Cite this article: Case BKM, Young J-G, Hébert-Dufresne L. 2023 Accurately summarizing an outbreak using epidemiological models takes time. *R. Soc. Open Sci.* **10**: 230634. https://doi.org/10.1098/rsos.230634

Received: 10 May 2023
Accepted: 30 August 2023

Subject Category:
Mathematics

Subject Areas:
health and disease and epidemiology/mathematical modelling/computational biology

Keywords:
practical identifiability, epidemiological modelling, Bayesian statistics

Author for correspondence:
Laurent Hébert-Dufresne
e-mail: laurent.hebert-dufresne@uvm.edu

Downloaded from https://royalsocietypublishing.org/ on 17 October 2023

Accurately summarizing an outbreak using epidemiological models takes time

B. K. M. Case[1,2], Jean-Gabriel Young[1,2,3] and Laurent Hébert-Dufresne[1,2]

[1]Vermont Complex Systems Center, [2]Department of Computer Science, and [3]Department of Mathematics and Statistics, University of Vermont, Burlington, VT 05405, USA

BKMC, 0000-0002-5666-1161; J-GY, 0000-0002-4464-2692

Recent outbreaks of Mpox and Ebola, and worrying waves of COVID-19, influenza and respiratory syncytial virus, have all led to a sharp increase in the use of epidemiological models to estimate key epidemiological parameters. The feasibility of this estimation task is known as the practical identifiability (PI) problem. Here, we investigate the PI of eight commonly reported statistics of the classic susceptible–infectious–recovered model using a new measure that shows how much a researcher can expect to learn in a model-based Bayesian analysis of prevalence data. Our findings show that the basic reproductive number and final outbreak size are often poorly identified, with learning exceeding that of individual model parameters only in the early stages of an outbreak. The peak intensity, peak timing and initial growth rate are better identified, being in expectation over 20 times more probable having seen the data by the time the underlying outbreak peaks. We then test PI for a variety of true parameter combinations and find that PI is especially problematic in slow-growing or less-severe outbreaks. These results add to the growing body of literature questioning the reliability of inferences from epidemiological models when limited data are available.

1. Introduction

Incredible efforts have been made in recent years to apply epidemiological models to the empirical data borne out of the COVID-19 pandemic. The LitCovid aggregator currently contains over 3000 papers on 'epidemic forecasting' and 'modelling and estimating' trends of COVID-19 spread [1]. We are seeing similar waves of models and forecasts for recent outbreaks of Mpox, Ebola, influenza and respiratory syncytial

© 2023 The Authors. Published by the Royal Society under the terms of the Creative Commons Attribution License http://creativecommons.org/licenses/by/4.0/, which permits unrestricted use, provided the original author and source are credited.

THE ROYAL SOCIETY PUBLISHING

Downloaded from https://royalsocietypublishing.org/ on 17 October 2023

virus. However, the enormous variability in model predictions, even among works using the same model and similar data, erodes confidence when interpreting these efforts for policy decisions [2]. It is clear that uncertainty remains about what we can expect to learn from these models, and when.

Disease models tackle the difficult challenge of describing complex epidemic processes by relating mechanistic processes to population-level observations such as daily reported cases. Identifying combinations of parameters that plausibly replicate observed data can help summarize the epidemic dynamics. Common statistics include the basic reproductive number, the average number of new cases someone will cause in an entirely susceptible population, and the outbreak size, the fraction of the population who will eventually have had the disease. Because these indicators are the product of interacting social and biological phenomena, they are never available through direct observation. Fitting epidemiological models to data is one of the best options for estimating these important quantities [3].

The classic susceptible–infectious–recovered (SIR) model accounts for a minimal number of critical mechanisms of disease spread. Infectious individuals infect susceptible individuals at a rate β and recover at a rate α. These mechanisms can be tracked through time by a set of ordinary differential equations:

$$\frac{\mathrm{d}}{\mathrm{d}t}S = -\beta SI, \quad \frac{\mathrm{d}}{\mathrm{d}t}I = \beta SI - \alpha I \quad \text{and} \quad \frac{\mathrm{d}}{\mathrm{d}t}R = \alpha I.$$

It is common to consider S, I and R as a fraction of the population in a given state such that $S + I + R = 1$ at all times. The initial state of the population might not be known—especially the susceptible pool $S_0 \equiv S(t = 0)$. Focusing on the second equation, we can see that the epidemic will grow exponentially at a rate $\beta S_0 - \alpha$ for initial small values of I, resulting in near-exchangeability of the parameters and causing large uncertainty in individual parameter values early on [4,5]. Conversely, when I becomes small after the peak, the infectious population eventually decays exponentially at a rate α. These observations make clear that data regarding I will provide information about different parameters, or combinations thereof, at different points of an outbreak. In general, the amount of information that can be learned about a given quantity will depend on the structure of the model equations, the timing of observations and the level of noise in the data [6].

Despite the model's simplicity, several authors have cautioned that the reliability of inferences drawn from the SIR model is questionable when based on prevalence data alone [7]. Due to the structural nature of the SIR equations, these issues are particularly acute during the early stage of an outbreak, when inferences are critical for informing timely public health response [5,8]. Without careful incorporation of additional data, these reliability problems can only grow with additional complexity in the model equations or observational structure [2,9]. In order to draw meaningful conclusions, researchers are forced to rely on data from one or more epidemic waves [10], or make strong and potentially controversial assumptions about parameters governing disease spread [11]. A more general understanding of how properties of epidemiological models affect uncertainty in commonly reported summary statistics would help researchers quantify how much they can expect to learn in empirical studies and establish sufficient criteria for reproducibility. Therefore, the goal of this report is to provide a comprehensive baseline for the reliability of estimates for a number of commonly reported statistics, with emphasis on the time necessary to predict these statistics in an emerging epidemic accurately and to illuminate the structural interactions between data, model dynamics and summary statistics.

This question of whether quantities estimated from data are reliable, e.g. compatible with some hypothetical true parameters $\boldsymbol{\theta}^* = (\alpha^*, \beta^*, S_0^*)$ which generated the data, is termed the practical identifiability (PI) problem and has traditionally been studied using the variance–covariance matrix of an estimator for $\boldsymbol{\theta}^*$ [12]. However, such second-order approaches underestimate uncertainty in limited data settings, where the distribution of plausible parameters may be skewed [13,14]. Here we propose a new measure that allows us to efficiently and directly measure our ability to learn various epidemiological quantities at all stages of an epidemic. If $u = f(\boldsymbol{\theta})$ is an unknown variable to be estimated, our Bayesian interpretation of the identifiability of u is the expected logarithm of the ratio between posterior and prior probabilities, evaluated at $u^* = f(\boldsymbol{\theta}^*)$:

$$\delta_u(\boldsymbol{\theta}^*) = \mathbf{E}_{\mathbf{y}|\boldsymbol{\theta}^*}[\log P(u^* \mid \mathbf{y}) - \log P(u^*)], \tag{1.1}$$

where $\mathbf{y} \mid \boldsymbol{\theta}^*$ are noisy observations of the underlying outbreak, e.g. daily case counts, and where the expectation is taken over realizations of the observation process. Since shrinkage in the posterior distribution is facilitated through the global behaviour of the model likelihood, (1.1) is able to capture uncertainty arising from complex model fits, such as bimodality in the likelihood surface. As with traditional approaches to PI, δ_u is a local measure of information gain, in the sense that changing the

Downloaded from https://royalsocietypublishing.org/ on 17 October 2023

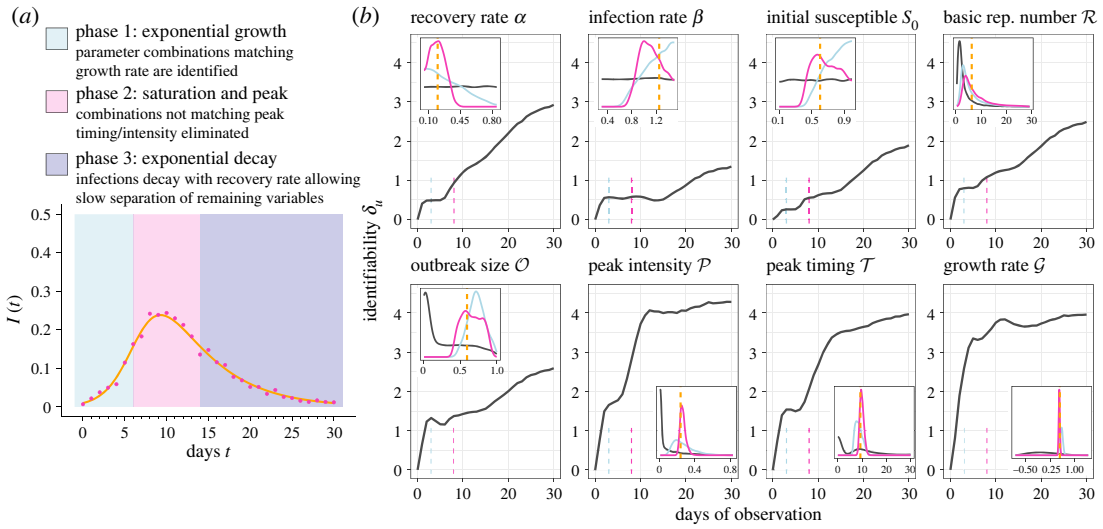

Figure 1. Practical identifiability of epidemiological summary statistics over time. (*a*) Unknown deterministic SIR process based on true parameters $\boldsymbol{\theta}^*$ (orange line), and single realization of observed data $\mathbf{y} \sim P(\mathbf{y} \mid \boldsymbol{\theta}^*)$ (pink dots). (*b*) Main panels show PI according to δ_u over an increasing observation window assuming daily observations. Insets give an example of how δ_u is interpreted, showing $P(u|\mathbf{y})$ and $P(u)$ for the single realization of $\mathbf{y}$ from (*a*), observed up to $T = 3$ (blue) and $T = 8$ (pink). The dashed orange line is the true value to be estimated. True parameters are $\alpha^* = 0.2$, $\beta^* = 1.25$ and $S_0^* = 0.6$, with $I_0 = 0.01$ assumed known. Prior beliefs are $\alpha \sim U(0.05, 0.85)$, $\beta \sim U(0.3, 1.5)$, $S_0 \sim U(0.1, 0.99)$.

Table 1. Definitions of epidemiological summary statistics.

name	symbol	formula
reproductive number	$\mathcal{R}$	β/α
outbreak size	$\mathcal{O}$	$1 - R(0) - S_0 \exp(-\mathcal{R}\mathcal{O})$[a]
peak intensity	$\mathcal{P}$	$I_0 + S_0 + [1 - \log(S_0/\mathcal{R})]/\mathcal{R}$
peak timing	$\mathcal{T}$	unknown
growth rate	$\mathcal{G}$	$\beta S_0 - \alpha$

[a]Implicit equation.

true dynamics $\boldsymbol{\theta}^*$ will, in general, give different answers [15]. This allows the effect of particular values of $\boldsymbol{\theta}^*$ to be studied. Note that the metric does not require computationally expensive Bayesian inference methods to compute—a simple Monte Carlo procedure for estimating (1.1) is provided in appendix A.

2. Results

Figure 1 shows the PI of the SIR model parameters, as well as five summary variables which are commonly calculated in terms of $\boldsymbol{\theta}$ (see table 1 for mathematical definitions), for a typical parametrization $\boldsymbol{\theta}^*$ of the model. Infectious individuals are assumed to be independently tested at a fixed rate η at daily timepoints, giving a likelihood $y_t \sim \text{Poisson}(\eta I(t; \boldsymbol{\theta}^*))$. We assumed $\eta = 1000$ is known throughout, which leads to limited observational noise to better study PI inherent to the SIR equations. δ_u is computed daily for the eight variables, up to a maximum of 30 days of observation.

The rate of learning for all variables is uneven over time, with each reaching plateaus of varying length before the peak. The infection rate β is the worst identified. Gaining information on α appears easier than β and S_0 and even exceeds learning for $\mathcal{R}$ and $\mathcal{O}$ after around $T = 20$ days of observation. PI of the peak intensity, peak timing and growth rate increase more rapidly at first, with learning for growth rate happening particularly fast.

These findings illustrate the difficulty of learning key quantities early in an epidemic, under real-time conditions where the number of observations increases as the outbreak goes on. However, the question remains as to what extent a lack of early learning may be attributed simply to a smaller sample size.

Downloaded from https://royalsocietypublishing.org/ on 17 October 2023

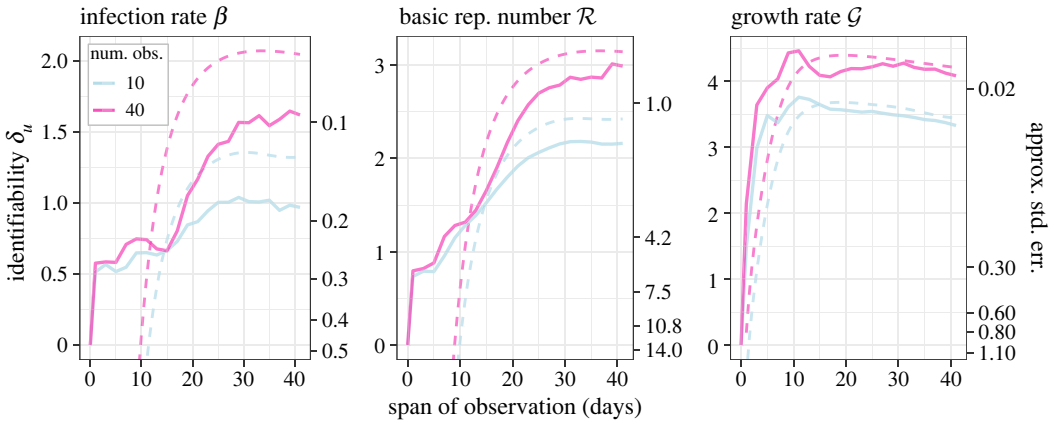

Figure 2. Practical identifiability of several variables as a function of testing frequency. Observations are evenly distributed over the interval [0, T] for increasing days of observation T. Solid lines are PI calculated using Monte Carlo, while dashed lines indicate the approximation given by (B 3). The approximation also gives an asymptotic relationship between δ_u and a lower bound on the standard error, indicated with secondary axes. Priors and true parameters are the same as in figure 1.

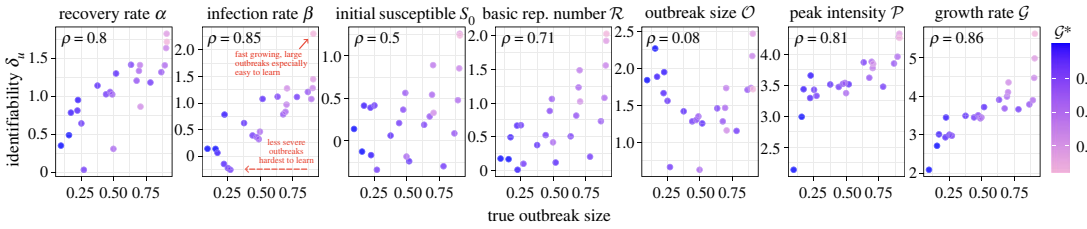

Figure 3. Change in PI δ_u when the true parameters $\boldsymbol{\theta}^*$ are varied. δ_u is calculated using daily observations up to the first day after the true (unobserved) outbreak has peaked. Each dot corresponds to a combination of true parameters with $\beta^* = 0.3, 0.5, \ldots, 1.5$ and $S_0^* = 0.1, 0.3, \ldots, 0.9$ while $\alpha^* = 0.2$ is fixed. Pearson correlation between δ_u and true outbreak size is given in the corners of each panel. Priors are the same as in figure 1.

Therefore, we next examined the PI of several variables over an increasing observation window, but with the number of evenly distributed observations kept constant. Figure 2 shows that identifiability of β and $\mathcal{R}$ is lowest when observations are concentrated prior to the peak, confirming that the limits of early learning are indeed a structural property of the SIR equations that cannot be overcome by allocating additional tests early on. Further, increasing the frequency of testing from 10 observations to 40 did little to increase PI during this period, but increased PI considerably for wider observation windows. Figure 2 also shows the functional relationship between the asymptotic limit of δ_u and the usual standard error for u, as given by (B 3), which can serve as an alternative interpretation of δ_u when there are sufficient data.

To test the sensitivity of these findings to $\boldsymbol{\theta}^*$, we then computed δ_u over a grid of values for β^* and S_0^* (figure 3). Since slower-growing outbreaks will naturally contain less information per day [7], information gain was calculated using observations up until the first day after the epidemic peak. To investigate the factors of a true outbreak most associated with learning, for each true value of the eight variables considered, the correlation between δ_u for each variable and the true value was computed. The outbreak size of the true epidemic was the most correlated with learning, followed by true growth rate, illustrating that less-severe outbreaks are harder to learn.

3. Discussion

The analysis presented here makes it clear that some epidemiological variables are easier to estimate through model dynamics than others, and emphasizes that most epidemiological summary statistics should be interpreted with caution when data are limited. Taken together, the rate of learning for all the variables suggests that learning takes place in three general phases. In Phase 1, plausible parameter combinations quickly concentrate along the surface $\{\boldsymbol{\theta}: \beta S_0 - \alpha = \mathcal{G}^*\}$, as infections increase exponentially with the initial growth rate. This explains the sharp but modest gain in the information

Downloaded from https://royalsocietypublishing.org/ on 17 October 2023

of all variables except for $\mathcal{G}$ during this phase. In Phase 2, infections begin to saturate and parameter combinations matching the true peak intensity and timing become more plausible. However, for β especially, saturating case counts do little to further restrict the plausible parameter surface from Phase 1. Finally, Phase 3 is characterized by gradual information gain for the remaining variables. Since infections are slowly declining with α during this phase, this growth is explained by α^* gradually being identified, which propagates to allow some remaining combinations on the plausible surface to be eliminated.

Parameters describing the mechanisms of the model—β, α and S_0—take a particularly long time to learn on account of quickly reaching a plateau at low values of δ_u. As a result, the SIR model is more effective at forecasting short-term statistics of the dynamics, such as peak timing and intensity, than it is at estimating mechanisms. This result shows how difficult it is to estimate parameters from early data in the hope of forecasting the impacts of mechanistic interventions such as reducing β with preventive measures or increasing α with treatment [16]. Importantly, even though a lack of identifiability implies a wide range of parameters lead to similar infectious dynamics early on, these plausible dynamics will still respond differently to interventions targeting specific mechanisms [17]. Thus, low PI simply means that an intervention's impact is difficult to forecast ahead of time.

Learning was nearly as difficult for the statistics $\mathcal{R}$ and $\mathcal{O}$ as for the individual model parameters, despite the fact that optimistically, these transformations would combine the information of each parameter they depend on. The failure of these statistics to resolve closely exchangeable parameter combinations limits their reliability for succinctly describing an epidemic. By contrast, the initial growth rate resolves such combinations to give rapid shrinkage to the correct value, despite encoding similar information to $\mathcal{R}$ about disease dynamics [18]. This suggests growth rates are a more reliable 'first look' at an outbreak when using prevalence data under the SIR model.

When varying the true values $\boldsymbol{\theta}^*$, see figure 3, we find that less-severe outbreaks are generally harder to learn, despite having more daily observations available before their peak. The initial susceptible population S_0 appears the most poorly identified across values of $\boldsymbol{\theta}^*$ by the peak, and the expected posterior shrinkage is even slightly negative for 25% of the tested values. An interesting implication for control measures is that the more we reduce the severity of true infection dynamics, the harder it will be to accurately estimate the impacts of interventions. Further, the mode of intervention matters: variability along the y-axis in figure 3 for similar values of $\mathcal{O}^*$ shows lowering S_0^* impacts learning differently than a reduction in β^*.

Previous investigations into the PI of the SIR model have mainly focused on the PI of α and β under the simplified model where $S_0 \approx 1$ is known. These works generally agree that PI of both α and β is limited during Phase 1 [4,5], but that the majority of information available has been learned by the time the disease has peaked [9,19,20]. Most comparably to the observational design in figure 1, Capaldi *et al.* [7] considered the asymptotic variance of $\hat{\beta}$ and $\hat{\alpha}$ over an increasing timespan, and found the variance of both estimators decreased rapidly and smoothly just before and after the peak, respectively [7]. In contrast, the uneven rate of learning of these parameters in figure 1 paints a more nuanced and pessimistic picture of PI when exact likelihoods and prior context are taken into account. This finding supports the idea that previous PI results based on approximation theory underestimate uncertainty, particularly during the early stages of an outbreak when the likelihood surface is highly nonlinear [13,21].

In this work, we have proposed a novel means of assessing PI which measures the expected posterior gain in density at the true value u^*. While comparing densities at a specific value may seem to ignore uncertainty in the posterior as a whole, we argue that δ_u is better interpreted as a measure of *shrinkage* rather than density, by marginalizing the global curvature of the likelihood onto a single dimension for u. If the projected span of high likelihood values is more narrow than the support of $P(u)$, shrinkage will occur and δ_u becomes positive. In this sense, (1.1) might be viewed as a quantitative alternative to the popular profile likelihood method, in which potential plateaus in the likelihood surface are projected to the space of some parameter θ_i and examined graphically [22]. Additionally, as shown in appendix B, δ_u may be interpreted in terms of standard measurements of uncertainty—in the limit of large data and under certain conditions, δ_u converges to a density form of the usual standard error of the maximum likelihood estimator, penalized by the prior weight. Therefore, while our measure was specifically designed to give a more accurate picture of uncertainty in limited data regimes, it also has asymptotic behaviour similar to the coefficient of variation for u.

The Bayesian nature of our method of assessing PI means that estimates of model parameters and any variables which depend on them are sensitive to prior beliefs. In this report, our choice of uniform priors represents modest assumptions about an emerging pathogen: *a priori*, just over 50% of scenarios result in

Downloaded from https://royalsocietypublishing.org/ on 17 October 2023

an outbreak (i.e. have $\beta S_0/\alpha > 1$), and outbreaks range from modest to highly severe (70% of individuals infected at peak). However, for many pathogens, more informative prior information is frequently available, for example on the recovery rate of a disease [23]. Relative to more realistic prior settings, this may mean α is more difficult to gain information about than β and S_0.

While we have considered only noisy observation of the current infectious population, real data may also come in the form of daily new infections or cumulative case counts, and may suffer from lags in reporting or preferential sampling [24,25]. Learning epidemiological variables from such data will have their own distinct challenges [9]. PI of the SIR model should also be assessed with hierarchical models incorporating data from multiple sources, such as hospitalizations and isolated clinical experiments [26]. Yet, our work shows that even in its simplest form, learning parameters and statistics of SIR dynamics takes time, limiting which inferences, forecasts and control policies can be made from early epidemic data.

Ethics. This work did not require ethical approval from a human subject or animal welfare committee.

Data accessibility. Materials necessary to reproduce this analysis are available on GitHub and have been archived on Zenodo [27]. A Julia package MarginalDivergence.jl has been developed for efficient computation of our PI measure, including an interface for easy implementation of user-provided models. The package is currently unregistered.

Declaration of AI use. We have not used AI-assisted technologies in creating this article.

Authors' contributions. B.K.M.C.: conceptualization, formal analysis, investigation, methodology, software, visualization, writing—original draft, writing—review and editing; J.-G.Y.: conceptualization, funding acquisition, investigation, methodology, supervision, writing—original draft, writing—review and editing; L.H.-D.: conceptualization, funding acquisition, investigation, methodology, supervision, writing—original draft, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. We declare we have no competing interests.

Funding. B.K.M.C., J.-G.Y. and L.H.-D. acknowledge support from the National Institutes of Health 1P20 GM125498-01 Centers of Biomedical Research Excellence Award. B.K.M.C. is also supported as a Fellow of the National Science Foundation under NRT award no. DGE-1735316, and L.H.-D. by the National Science Foundation award no. EPS-2019470.

Appendix A. Supplemental methods

A.1. Likelihood-based estimation of dynamical systems

While the methods considered here can be applied to any statistical process for which a likelihood exists, we are interested in processes of the form

$$y_i \sim g(\mathbf{x}(t_i), \boldsymbol{\sigma}) \tag{A 1}$$

and

$$\dot{\mathbf{x}} = h(\mathbf{x}(t), \boldsymbol{\tau}) \tag{A 2}$$

where $\mathbf{y} = (y_1, \ldots, y_n)$ are observations at discrete timepoints $t_1, \ldots, t_n$, and $\boldsymbol{\sigma}$, $\boldsymbol{\tau}$ are parameters that are assumed known or are to be estimated. We refer to h as the *latent process* and g as the *observation process*. We are interested in our ability to estimate a set of unknown parameters $\boldsymbol{\theta}^* \subseteq (\boldsymbol{\sigma}^*, \boldsymbol{\tau}^*, \mathbf{x}(\mathbf{0})^*)$.

Given $\boldsymbol{\theta}$, equations (A 1)–(A 2) form a probability distribution $P(\mathbf{y} \mid \boldsymbol{\theta})$ called the *likelihood*. In the frequentist paradigm, an estimator for $\boldsymbol{\theta}^*$ can be obtained by maximizing $P(\mathbf{y} \mid \boldsymbol{\theta})$:

$$\hat{\boldsymbol{\theta}}_{\mathrm{MLE}} = \mathrm{argmax}_{\boldsymbol{\theta}} P(\mathbf{y} \mid \boldsymbol{\theta}).$$

A popular way to assess issues of PI is through the variance–covariance matrix of $\hat{\boldsymbol{\theta}}_{\mathrm{MLE}}$, which can show marginal uncertainty in individual parameter estimators and correlations between pairs of estimators. The Cramer–Rao bound implies that in the limit of decreasing observation uncertainty (i.e. as the amount or precision of data increases), the variance of an unbiased estimator converges, given certain regularity conditions, to the inverse of the Fisher information matrix $\mathcal{I}(\boldsymbol{\theta}^*)$, where

$$[\mathcal{I}(\boldsymbol{\theta})]_{ij} = -\mathbf{E}_{\mathbf{y}|\boldsymbol{\theta}}\left[\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log P(\mathbf{y} \mid \boldsymbol{\theta})\right]. \tag{A 3}$$

This bound can underestimate variance when measurement noise is not infinitesimal [14,21], leading some to question its applicability even for simple nonlinear models [13,28]. An alternative is to estimate the distribution of $\hat{\boldsymbol{\theta}}_{\mathrm{MLE}}$ using Monte Carlo simulation, by sampling possible datasets $\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \ldots$ from

Downloaded from https://royalsocietypublishing.org/ on 17 October 2023

$P(\mathbf{y} \mid \boldsymbol{\theta}^*)$ and finding the maximum of each likelihood $P(\mathbf{y}^{(j)} \mid \boldsymbol{\theta})$ using an optimization algorithm such as gradient descent. The resulting samples $\hat{\boldsymbol{\theta}}^{(j)}_{\mathrm{MLE}}$ can then be inspected graphically or used to estimate the covariance matrix. This method has the convenience of also working with estimates of transformations of the model parameters, without the need for further approximation [19,29].

A.2. Proposed method of assessing practical identifiability

While using Monte Carlo estimation of $\mathrm{Var}(\hat{\boldsymbol{\theta}}_{\mathrm{MLE}})$ to assess PI can alleviate the underestimation issues when using the information matrix, the use of optimization to obtain a sample of the estimator can lead to dependence on initial conditions or other hyperparameters of the optimization method used [12]. Again, the inaccuracy of this method will be most acute when the likelihood surface is flat or multi-modal, such as when limited data are available.

Rather than relying on optimization, we instead take a sampling-based Bayesian perspective. From the main text, we have for a variable of interest $u = f(\boldsymbol{\theta})$, $\delta_u(\boldsymbol{\theta}^*) = \mathbf{E}_{\mathbf{y}|\boldsymbol{\theta}^*}[\log P(u^* \mid \mathbf{y})] - \log P(u^*)$, which gives the average amount, over possible future outbreaks $P(\mathbf{y} \mid \boldsymbol{\theta}^*)$, a researcher can expect to learn about the true quantity u^* in a Bayesian analysis. A value of $\delta_u = c$ corresponds roughly to an expected gain in posterior probability c orders of magnitude greater than the prior.

Equation (1.1) can be rewritten by applying Bayes' rule, $P(u^* \mid \mathbf{y})/P(u^*) = P(\mathbf{y} \mid u^*)/P(\mathbf{y})$, where the margin $P(\mathbf{y} \mid u^*)$ equals $\int P(\mathbf{y} \mid \boldsymbol{\theta}) P(\boldsymbol{\theta} \mid u^*) \mathrm{d}\boldsymbol{\theta}$ and $P(\boldsymbol{\theta} \mid u^*)$ is the distribution of the epidemiological parameters compatible with a fixed variable of interest u^*—we give details below. This leads to

$$\delta_u(\boldsymbol{\theta}^*) = \mathbf{E}_{\mathbf{y}|\boldsymbol{\theta}^*}\left[\log \frac{P(\mathbf{y} \mid u^*)}{P(\mathbf{y})}\right]. \tag{A 4}$$

We approximate $\delta_u(\boldsymbol{\theta}^*)$ by generating M paired Monte Carlo samples from $P(\boldsymbol{\theta} \mid u^*)$ and $P(\boldsymbol{\theta})$, and reusing these samples to obtain M samples from $P(\mathbf{y} \mid u^*)$ and $P(\mathbf{y})$ for each $\mathbf{y} \sim P(\mathbf{y} \mid \boldsymbol{\theta}^*)$, leading to

$$\delta_u(\boldsymbol{\theta}^*) \approx \frac{1}{N} \sum_{i=1}^{N} \left[\log \sum_{j=1}^{M} P(\mathbf{y}^{(i)} \mid \tilde{\boldsymbol{\theta}}^{(j)}) - \log \sum_{j=1}^{M} P(\mathbf{y}^{(i)} \mid \boldsymbol{\theta}^{(j)})\right]$$

where $\tilde{\boldsymbol{\theta}}^{(j)} \sim P(\boldsymbol{\theta} \mid u^*)$, $\boldsymbol{\theta}^{(j)} \sim \boldsymbol{P}(\boldsymbol{\theta})$ and $\mathbf{y}^{(i)} \sim P(\mathbf{y} \mid \boldsymbol{\theta}^*)$. $N = 3000$ and $M = 60\,000$ were used for all computations in the main text.

A.2.1. Accuracy of Monte Carlo estimation of δ_u

The marginal likelihood $P(\mathbf{y})$ is notorious for being inefficient to estimate via Monte Carlo methods. To test our choice of M was large enough while still within a reasonable computational budget, we repeated calculations of $\log P(\mathbf{y})$ for increasing values of M, where a $\mathbf{y} \sim P(\mathbf{y} \mid \boldsymbol{\theta}^*)$ was sampled with 60 observations (every half day). $\boldsymbol{\theta}^*$ and $P(\boldsymbol{\theta})$ were the same as in figure 1 of the main text. We concluded that even with 60 observations, which gives a likelihood sharper than the maximum 30 observations used in the main text, a choice of $M > 30\,000$ was sufficient to give a standard error less than 1, or less than 0.5% of the magnitude of $\log P(\mathbf{y})$. The runtime and standard errors from 100 independent computations of $P(\mathbf{y})$ are shown as a function of M in figure 4.

A.3. Practical identifiability for a function of model parameters

The distribution function $P(\boldsymbol{\theta} \mid u^*)$ will generally not be available in closed form even when $P(\boldsymbol{\theta})$ is.

Simulating from $P(\boldsymbol{\theta} \mid u^*)$ can be accomplished with the following procedure: let $\theta_i \in \boldsymbol{\theta}$ be a chosen 'pivot' parameter/index and define $\tilde{f}(\theta_i \mid \boldsymbol{\theta}_{-i}) = u$ to be a univariate function conditional on $\boldsymbol{\theta}_{-i}$, where $\boldsymbol{\theta}_{-i}$ indicates the ith element of $\boldsymbol{\theta}$ has been removed. Assume $\tilde{f}$ is invertible so that $\tilde{f}^{-1}(u \mid \boldsymbol{\theta}_{-i}) = \theta_i$. Then, assuming independent priors on the elements of $\boldsymbol{\theta}$, using a change of variables and Bayes' rule we have

$$P(\boldsymbol{\theta}_{-i} \mid u^*) \propto \prod_{j \neq i} P(\theta_j) P(u^* \mid \boldsymbol{\theta}_{-i}), \tag{A 5}$$

where

$$P(u \mid \boldsymbol{\theta}_{-i}) = \left|\frac{\mathrm{d}}{\mathrm{d}u} \tilde{f}^{-1}(u \mid \boldsymbol{\theta}_{-i})\right| P_{\theta_i}(\tilde{f}^{-1}(u^* \mid \boldsymbol{\theta}_{-i})). \tag{A 6}$$

Because $\tilde{f}$ is deterministic (i.e. θ_i can be uniquely determined given $\boldsymbol{\theta}_{-i}$ and u), samples from $P(\boldsymbol{\theta} \mid u^*)$ can, therefore, be obtained by first sampling $\boldsymbol{\theta}^{(1)}_{-i}, \ldots, \boldsymbol{\theta}^{(n)}_{-i}$ from (A 5) using a standard simulation technique

Downloaded from https://royalsocietypublishing.org/ on 17 October 2023

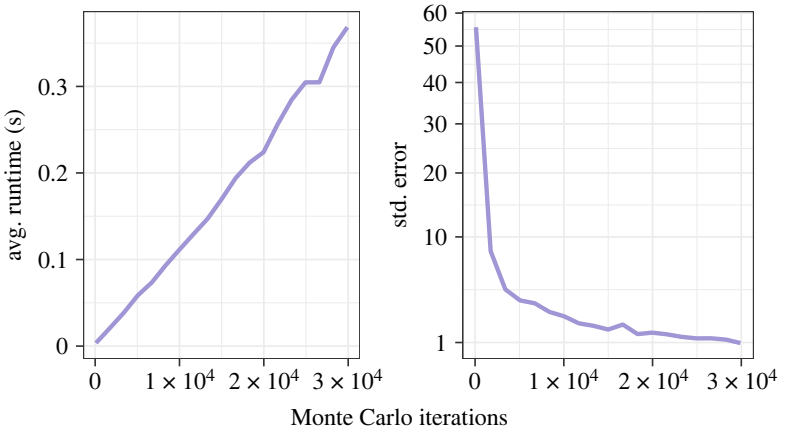

Figure 4. Speed and accuracy of approximating the log marginal likelihood log $P(\mathbf{y})$ using M Monte Carlo simulations. The results are averaged over 100 repetitions of the sampling process for each M.

such as accept–reject sampling, and then letting $\theta_i^{(j)} = \tilde{f}^{-1}(u^*, \boldsymbol{\theta}_{-i}^{(j)})$. The resulting densities for the five transformations considered here are shown in figure 5.

For example, under the transformation $f(\boldsymbol{\theta}) = \beta/\alpha = :\mathcal{R}$, we define $\tilde{f}^{-1}(\alpha, S_0, \mathcal{R}) = \alpha\mathcal{R}$ and obtain

$$P(\alpha, S_0 \mid \mathcal{R}) \propto P_\alpha(\alpha) P_{S_0}(S_0) \alpha P_\beta(\mathcal{R}\alpha). \tag{A 7}$$

So we may sample $(\alpha^{(1)}, S_0^{(1)}), (\alpha^{(2)}, S_0^{(2)}), \ldots$ from (A 7), then let $\beta^{(i)} = \mathcal{R}^* \alpha^{(i)}$ to obtain a sample from $P(\alpha, \beta, S_0 \mid \mathcal{R}^*)$.

For the final outbreak size, we define $\mathcal{O} := R(\infty) - R(0)$ to be the total proportion of individuals who end up in the recovered compartment due to infection. For $R(\infty)$, we have from Weiss [30]

$$R(\infty) = 1 - S_0 \exp(-\mathcal{R}(R(\infty) - R(0))), \tag{A 8}$$

which we may use to solve for β and obtain the inverse function

$$\beta = \frac{-\alpha}{\mathcal{O}} \log \frac{1 - R(0) - \mathcal{O}}{S_0}, \tag{A 9}$$

and the derivative

$$\frac{\mathrm{d}\beta}{\mathrm{d}\mathcal{O}} = \frac{\alpha}{\mathcal{O}} \left(\frac{1}{\mathcal{O}} \log \frac{1 - R(0) - \mathcal{O}}{S_0} + \frac{1}{1 - R(0) - \mathcal{O}} \right). \tag{A 10}$$

For the peak intensity $\mathcal{P} := \max_t I(t)$, to obtain samples from (A 5) we may use the equation

$$\mathcal{P} = I_0 + S_0 - \frac{\alpha}{\beta} \log S_0 - \frac{\alpha}{\beta} \left(1 + \log \frac{\alpha}{\beta} \right). \tag{A 11}$$

Although (A 11) yields only implicit solutions for any θ_i, a closed-form solution for S_0 given $\mathcal{P}$ can be found using Lambert's W:

$$S_0 = -\mathcal{R}^{-1} W_{-1}(-B), \tag{A 12}$$

where $B = \exp(-\mathcal{R}(\mathcal{P} - I_0) - 1)$, and derivative

$$\frac{\mathrm{d}S_0}{\mathrm{d}\mathcal{P}} = \frac{1}{1 - B\,\mathrm{e}^{W_{-1}(-B)}}. \tag{A 13}$$

Derivation of the necessary equations for the initial growth rate $\mathcal{G} := \beta S_0 = \alpha$ is straightforward.

Finally, the peak timing $\mathcal{T}$ does not have a known closed-form solution. Though more time-consuming, we can still approximate (A 5) by using univariate constrained optimization to evaluate the unknown f^{-1}, and adjoint methods to obtain the corresponding derivative.

Downloaded from https://royalsocietypublishing.org/ on 17 October 2023

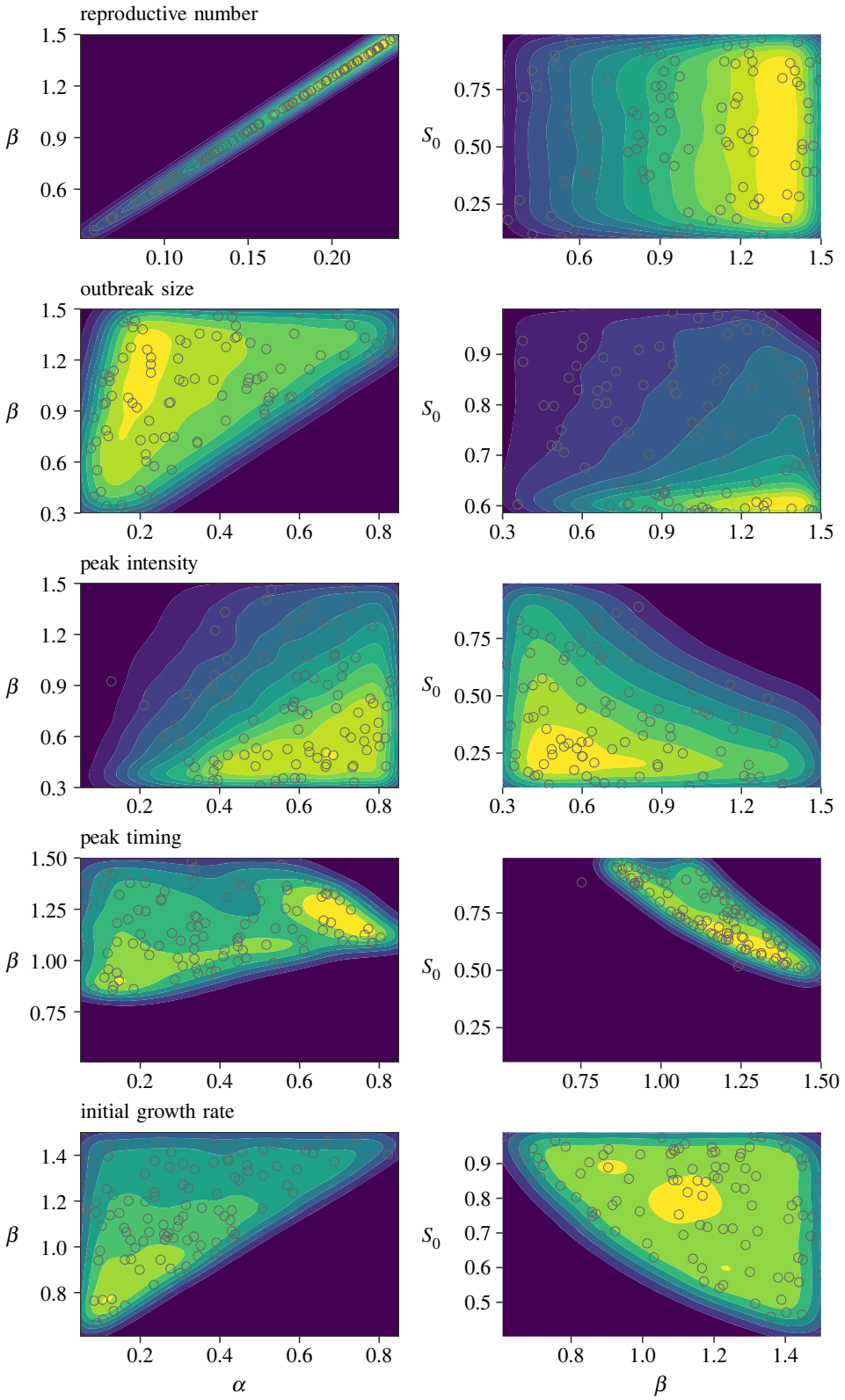

Figure 5. Density of 60 000 samples from $P(\alpha, \beta | u^*)$, and $P(\beta, S_0 | u^*)$ given different summary transformations. True values and priors are the same as in the main text.

Appendix B. Asymptotic properties of practical identifiability

The above procedure for estimating δ_u using simple Monte Carlo becomes inefficient when the dimension of $\boldsymbol{\theta}$ or $\mathbf{y}$ becomes large. In the latter case, the proposed method of PI can instead be analysed using the

Downloaded from https://royalsocietypublishing.org/ on 17 October 2023

usual approximation theory in the limit of large data. The Bernstein–von Mises theorem gives

$$P(\boldsymbol{\theta} \mid \mathbf{y}) \approx \mathcal{N}(\hat{\boldsymbol{\theta}}_{\mathrm{MLE}}, (n_t \mathcal{I}(\hat{\boldsymbol{\theta}}_{\mathrm{MLE}}))^{-1}), \tag{B 1}$$

where n_t is the number of independent replications of the time series of observations, and $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Under certain regularity conditions, $\hat{\boldsymbol{\theta}}_{\mathrm{MLE}} = \boldsymbol{\theta}^*$.

The information matrix of the model parameters can be separated in terms of the curvature of latent and observation processes. In the case where data from a single state variable x are observed, $\mathcal{I}(\boldsymbol{\theta}) = \mathbf{J}^\top \mathcal{O}(\boldsymbol{\theta})\mathbf{J}$, where $\mathbf{J}$ is the Jacobian of x with respect to $\boldsymbol{\theta}$, $\mathbf{J}_{ij} = \partial x(t_i)/\partial \boldsymbol{\theta}_j$, and $\mathcal{O}(\boldsymbol{\theta})$ is the information of $\mathbf{y}$ given x. In the case of independent Poisson-distributed testing, $\mathcal{O}(\boldsymbol{\theta}) = \mathrm{diag}(\eta I(t_1; \boldsymbol{\theta}), \ldots, \eta I(t_n; \boldsymbol{\theta}))$.

To obtain an approximation for the posterior of a transformation $u = f(\boldsymbol{\theta})$, we may again choose a pivot parameter/index θ_i and introduce a change of variables $\mathbf{v} = (\theta_1, \ldots, \theta_{i-1}, u, \theta_{i+1}, \ldots, \theta_p)$, and define the vector-valued function $\tilde{\mathbf{f}}$ so that $\tilde{\mathbf{f}}(\mathbf{v}) = \boldsymbol{\theta}$. Letting $\mathbf{F}$ be the gradient of $\tilde{\mathbf{f}}$ with respect to $\mathbf{v}$, $\mathbf{F}_{kj} = \partial\theta_k/\partial v_j$, we have

$$\mathcal{I}(\mathbf{v}) = \mathbf{F}^\top \mathbf{J}^\top \mathcal{O}(\boldsymbol{\theta})\mathbf{J}\mathbf{F}. \tag{B 2}$$

Combining (B 1) and (B 2), therefore, gives the approximation

$$\delta_u(\boldsymbol{\theta}^*) \approx \frac{1}{2}(\log n_t - \log \mathcal{I}(\mathbf{v}^*)_{ii}^{-1} - \log(2\pi)) - \log P(u^*). \tag{B 3}$$

This reveals that, in the limit of sufficient data, δ_u is related to the local curvature of the latent and observational processes, just like traditional asymptotic approaches to PI. For individual model parameters $u \in \boldsymbol{\theta}$, $\mathbf{F}$ is the identity matrix and δ_u becomes the logarithm of the usual standard error of the estimator $\hat{u}_{\mathrm{MLE}}$, penalized by the prior log-probability of $\hat{u}_{\mathrm{MLE}}$. For parameter transformations, the information of $\mathcal{I}(\boldsymbol{\theta})$ is then summarized further through the curvature induced by the transformation function f. Also note that penalizing by the prior density has a normalizing effect, as transformations that increase the support of $P(u)$ will also have smaller prior densities, and therefore is analogous to using the coefficient of variation to allow comparing standard errors between variables.

References

1. Chen Q, Allot A, Lu Z. 2021 LitCovid: an open database of COVID-19 literature. *Nucleic Acids Res.* **49**, D1534–D1540. (doi:10.1093/nar/gkaa952)
2. Roda WC, Varughese MB, Han D, Li MY. 2020 Why is it difficult to accurately predict the COVID-19 epidemic? *Infect. Dis. Model.* **5**, 271–281. (doi:10.1016/j.idm.2020.03.001)
3. Wu JT, Leung K, Leung GM. 2020 Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. *Lancet* **395**, 689–697. (doi:10.1016/S0140-6736(20)30260-9)
4. Chowell G. 2017 Fitting dynamic models to epidemic outbreaks with quantified uncertainty: a primer for parameter uncertainty, identifiability, and forecasts. *Infect. Dis. Model.* **2**, 379–398. (doi:10.1016/j.idm.2017.08.001)
5. Melikechi O, Young AL, Tang T, Bowman T, Dunson D, Johndrow J. 2022 Limits of epidemic prediction using SIR models. *J. Math. Biol.* **85**, 36. (doi:10.1007/s00285-022-01804-5)
6. Seber GAF, Wild CJ. 2003 *Nonlinear regression*, 2nd edn. Hoboken, NJ: John Wiley & Sons.
7. Capaldi A, Behrend S, Berman B, Smith J, Wright J, Lloyd A. 2012 Parameter estimation and uncertainty quantification for an epidemic model. *Math. Biosci. Eng.* **9**, 553–576. (doi:10.3934/mbe.2012.9.553)
8. Castro M, Ares S, Cuesta JA, Manrubia S. 2020 The turning point and end of an expanding epidemic cannot be precisely forecast. *Proc. Natl Acad. Sci. USA* **117**, 26 190–26 196. (doi:10.1073/pnas.2007868117)
9. Tuncer N, Le TT. 2018 Structural and practical identifiability analysis of outbreak models. *Math. Biosci.* **299**, 1–18. (doi:10.1016/j.mbs.2018.02.004)
10. Subramanian R, He Q, Pascual M. 2021 Quantifying asymptomatic infection and transmission of COVID-19 in New York City using observed cases, serology, and testing capacity. *Proc. Natl Acad. Sci. USA* **118**, e2019716118. (doi:10.1073/pnas.2019716118)
11. Britton T, Ball F, Trapman P. 2020 A mathematical model reveals the influence of population heterogeneity on herd immunity to SARS-CoV-2. *Science* **369**, 846–849. (doi:10.1126/science.abc6810)
12. Lam NN, Docherty PD, Murray R. 2022 Practical identifiability of parametrised models: a review of benefits and limitations of various approaches. *Math. Comput. Simul.* **199**, 202–216. (doi:10.1016/j.matcom.2022.03.020)
13. Joshi M, Seidel-Morgenstern A, Kremling A. 2006 Exploiting the bootstrap method for quantifying parameter confidence intervals in dynamical systems. *Metab. Eng.* **8**, 447–455. (doi:10.1016/j.ymben.2006.04.003)
14. Raman DV, Anderson J, Papachristodoulou A. 2017 Delineating parameter unidentifiabilities in complex models. *Phys. Rev. E* **95**, 032314. (doi:10.1103/PhysRevE.95.032314)
15. Sharp J, Browning A, Burrage K, Simpson MJ. 2022 Parameter estimation and uncertainty quantification using information geometry. *J. R. Soc. Interface* **19**, 20210940. (doi:10.1098/rsif.2021.0940)
16. Barnett M, Buchak G, Yannelis C. 2023 Epidemic responses under uncertainty. *Proc. Natl Acad. Sci. USA* **120**, e2208111120. (doi:10.1073/pnas.2208111120)
17. Alahmadi A *et al.* 2020 Influencing public health policy with data-informed mathematical models of infectious diseases: recent developments and new challenges. *Epidemics* **32**, 100393. (doi:10.1016/j.epidem.2020.100393)
18. Bettencourt LM, Ribeiro RM. 2008 Real time Bayesian estimation of the epidemic potential of emerging infectious diseases. *PLoS ONE* **3**, e2185. (doi:10.1371/journal.pone.0002185)

19. Roosa K, Chowell G. 2019 Assessing parameter identifiability in compartmental dynamic models using a computational approach: application to infectious disease transmission models. *Theor. Biol. Med. Model.* **16**, 1. (doi:10.1186/s12976-018-0097-6)
20. Piazzola C, Tamellini L, Tempone R. 2021 A note on tools for prediction under uncertainty and identifiability of SIR-like dynamical systems for epidemiology. *Math. Biosci.* **332**, 108514. (doi:10.1016/j.mbs.2020.108514)
21. Hines KE, Middendorf TR, Aldrich RW. 2014 Determination of parameter identifiability in nonlinear biophysical models: a Bayesian approach. *J. Gen. Physiol.* **143**, 401–416. (doi:10.1085/jgp.201311116)
22. Raue A, Kreutz C, Maiwald T, Bachmann J, Schilling M, Klingmüller U, Timmer J. 2009 Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics* **25**, 1923–1929. (doi:10.1093/bioinformatics/btp358)
23. Cevik M, Tate M, Lloyd O, Maraolo AE, Schafers J, Ho A. 2021 SARS-CoV-2, SARS-CoV, and MERS-CoV viral load dynamics, duration of viral shedding, and infectiousness: a systematic review and meta-analysis. *Lancet Microbe* **2**, e13–e22. (doi:10.1016/S2666-5247(20)30172-5)
24. Chiu WA, Ndeffo-Mbah ML. 2021 Using test positivity and reported case rates to estimate state-level COVID-19 prevalence and seroprevalence in the United States. *PLoS Comput. Biol.* **17**, e1009374. (doi:10.1371/journal.pcbi.1009374)
25. Wu SL *et al.* 2020 Substantial underestimation of SARS-CoV-2 infection in the United States. *Nat. Commun.* **11**, 4507. (doi:10.1038/s41467-020-18272-4)
26. De Angelis D, Presanis AM, Birrell PJ, Tomba GS, House T. 2015 Four key challenges in infectious disease modelling using data from multiple sources. *Epidemics* **10**, 83–87. (doi:10.1016/j.epidem.2014.09.004)
27. Case BKM. 2023 Code for: Brendandaisy/epi-summaries-over-time: RSOS Materials (Version v1). *Zenodo*. (doi:10.5281/zenodo.8280113)
28. Krausch N, Barz T, Sawatzki A, Gruber M, Kamel S, Neubauer P, Cruz Bournazou MN. 2019 Monte Carlo simulations for the analysis of non-linear parameter confidence intervals in optimal experimental design. *Front. Bioeng. Biotechnol.* **7**, 122. (doi:10.3389/fbioe.2019.00122)
29. Chowell G, Hengartner NW, Castillo-Chavez C, Fenimore PW, Hyman JM. 2004 The basic reproductive number of Ebola and the effects of public health measures: the cases of Congo and Uganda. *J. Theor. Biol.* **229**, 119–126. (doi:10.1016/j.jtbi.2004.03.006)
30. Weiss H. 2013 The SIR model and the foundations of public health. *Mater. Math.* **3**, 1–17.

Downloaded from https://royalsocietypublishing.org/ on 17 October 2023