scientific reports

OPEN

Identifiability of parameters in mathematical models of SARS-CoV-2 infections in humans

Stanca M. Ciupe[1] & Necibe Tuncer[2]

Determining accurate estimates for the characteristics of the severe acute respiratory syndrome coronavirus 2 in the upper and lower respiratory tracts, by fitting mathematical models to data, is made difficult by the lack of measurements early in the infection. To determine the sensitivity of the parameter estimates to the noise in the data, we developed a novel two-patch within-host mathematical model that considered the infection of both respiratory tracts and assumed that the viral load in the lower respiratory tract decays in a density dependent manner and investigated its ability to match population level data. We proposed several approaches that can improve practical identifiability of parameters, including an optimal experimental approach, and found that availability of viral data early in the infection is of essence for improving the accuracy of the estimates. Our findings can be useful for designing interventions.

Understanding the upper respiratory tract (URT) kinetics of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is important for designing public health interventions such as testing, isolation, quarantine, and drug therapies[1–12]. Similarly, understanding the kinetics of SARS-CoV-2 in the lower respiratory tract (LRT) is important for predicting the potential for severe disease, respiratory failure, and/or death[3,13]. Insights into the mechanism of SARS-CoV-2-host interactions and their role in transmission and disease have been found using mathematical models applied to longitudinal data[3–11,14–17]. While these studies are instrumental in determining important parameters (such as SARS-CoV-2 daily shedding and clearance rates, basic reproduction number, the role of innate immune responses in controlling and/or exacerbating the disease), their predictions are limited by the lack of data early in the infection. As such, with few (if any) samples available before viral titers peak, the early virus kinetics and the mechanisms for these early kinetics are uncertain. In this study, we investigate the sensitivity of the predicted outcomes of a within-host model of SARS-CoV-2 infection to the availability of data during different stages of the infection and use our findings to make recommendation.

A German study by Wolfel et al. collected data from nine patients infected early in the pandemic through contact with the same index case[18]. The study showed independent virus replication in upper and lower respiratory tracts[18,19] suggesting the possibility that virus kinetics, disease stages, and host involvement in control and pathogenesis are dependent on which area of the respiratory tract is homing SARS-CoV-2 at different stages of the disease[20–22]. One shortcoming when evaluating the data in this study comes from the fact that viral RNA was collected only after the patients became symptomatic, with an estimated first data point available on average 5–7 days after infection. Several within-host mathematical models developed and applied to the data set in the Wolfel et al. study have evaluated SARS-CoV-2 parameters, determined the role of innate immune responses, found connections between total RNA and infectious titers, and identified the efficacy of drug therapies[3,6,7]. We are interested in determining how the lack of data early in the infection affects these estimates.

We first developed our own within-host model that does not consider innate immunity explicitly. The model is an expansion of the within-host mathematical models developed for influenza and other respiratory infections[23–26] that includes both URT and LRT patches. We used the data from Wolfel et al. to estimate pertinent parameters and investigated the sensitivity of the estimated parameters to the presence of data at different stages of infection. To accomplish this, we created virtual data sets that span various stages of the infection and

[1]Department of Mathematics, Virginia Polytechnic Institute and State University, 225 Stanger Street, Blacksburg, VA 24060, USA. [2]Department of Mathematics, Florida Atlantic University, 777 Glades Road, Boca Raton, FL 33431, USA. ✉email: stanca@vt.edu

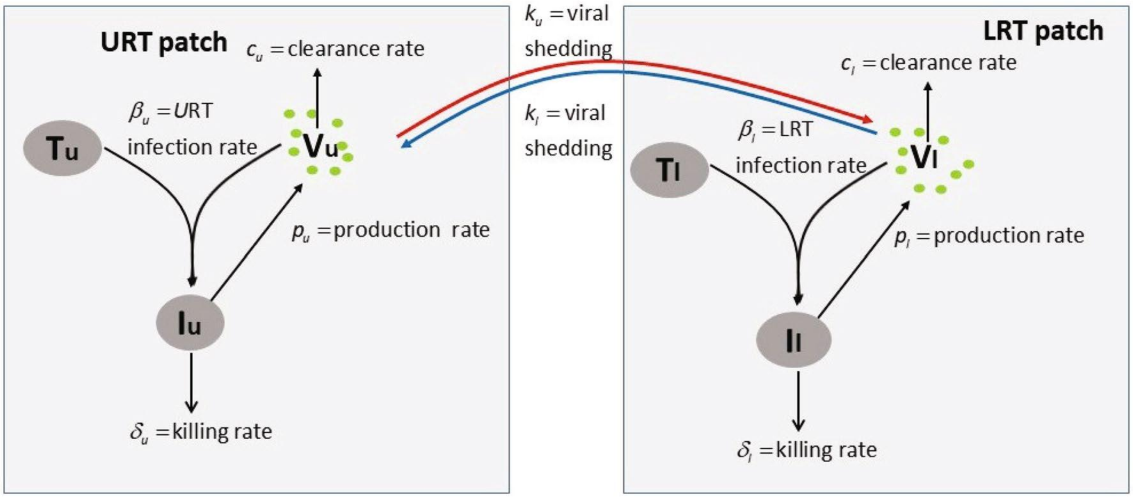

Figure 1. Model diagram.

determined how our initial predictions are being influenced by the additional data. Such results may influence our understanding of both viral expansion and the effect of inoculum dose on disease progression.

Methods

Mathematical model.

SARS-CoV-2 virus infects and replicates in epithelial cells of the upper and lower respiratory tract[18]. We model this by developing a two patch within-host model, where the patches are the two respiratory tracts which are linked through virus migration between patches, or viral shedding. Both respiratory tract patches assume interactions between uninfected epithelial cells, T_j; infected epithelial cells, I_j; and virus homing in tract j, V_j at time t. Here, $j = \{u, l\}$, with u describing the URT patch and l describing the LRT patch. Target cells in each patch get infected at rates β_j and infected cells produce new virions at rates p_j. Infected cells die at rates δ_j and virus particles are cleared at a linear rate c_u in the upper respiratory tract and in a density dependent manner $c_l V_l/(V_l + K)$ in the lower respiratory tract, where K is the viral load in the LRT where the clearance is half maximal. The two patches are linked via the virus populations, with a proportion k_u of V_u migrating from URT to LRT and k_l of V_l migrating from LRT to URT. The model describing these interactions (see Fig. 1) is given by

$$\begin{aligned}
\frac{dT_u}{dt} &= -\beta_u T_u V_u,\\
\frac{dI_u}{dt} &= \beta_u T_u V_u, -\delta_u I_u,\\
\frac{dV_u}{dt} &= p_u I_u - c_u V_u + k_l V_l,\\
\frac{dT_l}{dt} &= -\beta_l T_l V_l,\\
\frac{dI_l}{dt} &= \beta_l T_l V_l - \delta_l I_l,\\
\frac{dV_l}{dt} &= p_l I_l - c_l \frac{V_l}{V_l + K} V_l + k_u V_u.
\end{aligned} \tag{1}$$

We model the initial conditions of the system Eq. (1) as follows. We assume that all epithelial cells in the URT and LRT patches are susceptible to virus infection. When infection occurs, it results in a small initial virus inoculum which homes in the URT alone. Under these assumptions, system Eq. (1) is subject to initial conditions

$$\begin{aligned}
&T_u(0) = T_u^0,\ I_u(0) = 0,\ V_u(0) = V_0,\\
&T_l(0) = T_l^0,\ I_l(0) = 0,\ V_l(0) = 0,
\end{aligned} \tag{2}$$

where V_0 is the viral inoculum. We aim to determine the dynamics of system Eq. (1) over time for model parameters that explain URT and LRT tract data in a single patient (patient A) and in the population data (all nine patients) from[18].

Parameter estimation.

Patient data. In January 2020, nine patients tested positive for COVID-19 in a single-source outbreak in Bavaria, Germany[19]. Early detection allowed for rapid contact tracing, testing, and monitoring of the affected community: young healthy professionals in their mid-thirties. A followup study pub-

Patient id	A	B	C	D	E	F	G	H	I
Incubation period (days)	2.5	4	1	4	4	4	2	4.5	7

Table 1. Incubation periods estimated in Ref.[19].

lished time series for the post symptoms virus data isolated from oral-and nasopharyngeal throat swabs (in copies per swabs) and from sputum samples (in RNA copies per mL) for the same patient population over their entire course of disease. The patients' throat swabs and sputum data (Fig. 2 of [18]) were obtained through personal communication with the authors. Since we know the incubation period for each patient[19] (see Table 1), we assume time zero in our study to be the day of infection for the patients in Ref.[18].

Identifiability analysis. Using the URT and LRT viral load data, we aim to determine the unknown parameters $p = \{\beta_u, \delta_u, p_u, c_u, k_l, \beta_l, \delta_l, p_l, c_l, K, k_u\}$ of the within-host model Eq. (1). Before attempting to estimate the within-host model parameters using noisy laboratory data, it is crucial to analyze whether the model is structurally identifiable. Specifically, we need to know if the within-host model Eq. (1) is structured to reveal its parameters from upper and lower viral load observations. We approach this problem in an ideal setting where we assume that the observations are known for every $t > 0$ and they are not contaminated with any noise. This analysis is called structural identifiability[27].

The observed data in Wolfel et al.[18] is modeled in the within-host model Eq. (1) by variables V_u and V_l, which account for the upper and lower respiratory tract viral titers. We denote these observed variable as

$$y_1(t) = V_u(t) \quad \text{and} \quad y_2(t) = V_l(t).$$

First, we give the definition of structural identifiability in terms of the observed variables $y_1(t)$ and $y_2(t)$[27–31].

Definition 1 Let p and q be the two distinct vectors of within-host model Eq. (1) parameters. We say that the within-host model is structurally (globally) identifiable if and only if

$$y_1(t,p) = y_1(t,q) \quad \text{and} \quad y_2(t,p) = y_2(t,q) \quad \Longrightarrow \quad p = q.$$

Simply put, we say that the within-host model Eq. (1) is structurally identifiable if two identical observation are only possible for identical parameters. We perform the structural identifiability analysis via differential algebra approach. The first step in this approach is eliminating the unobserved state variables from the within-host model Eq. (1). The reason for eliminating the unobserved state variables is to obtain a system which only involves the observed states and model parameters. Since this is a complex procedure, we use DAISY[32] and obtain the following system

$$\begin{aligned}
&\frac{d^3y_1}{dt^3}y_1 - \frac{d^2y_1}{dt^2}\frac{dy_1}{dt} + \frac{d^2y_1}{dt^2}y_1^2\beta_u + \frac{d^2y_1}{dt^2}y_1(c_u+\delta_u) - \left(\frac{dy_1}{dt}\right)^2(c_u+\delta_u) + \frac{dy_1}{dt}\frac{dy_2}{dt}k_l \\
&\quad + \frac{dy_1}{dt}y_1^2\beta_u(c_u+\delta_u) + \frac{dy_1}{dt}y_2\delta_u k_l - \frac{d^2y_2}{dt^2}y_1k_l - \frac{dy_2}{dt}y_1^2\beta_u k_l - \frac{dy_2}{dt}y_1\delta_u k_l + y_1^3\beta_u c_u\delta_u - y_1^2y_2\beta_u\delta_u k_l = 0.
\end{aligned} \tag{3}$$

and

$$
\begin{aligned}
&-\frac{d^2y_1}{dt^2}y_2^5k_u-4\frac{d^2y_1}{dt^2}y_2^4Kk_u-6\frac{d^2y_1}{dt^2}y_2^3K^2k_u-4\frac{d^2y_1}{dt^2}y_2^2K^3k_u-\frac{d^2y_1}{dt^2}y_2K^4k_u+\frac{dy_1}{dt}\frac{dy_2}{dt}y_2^4k_u\\
&+4\frac{dy_1}{dt}\frac{dy_2}{dt}y_2^3Kk_u+6\frac{dy_1}{dt}\frac{dy_2}{dt}y_2^2K^2k_u+4\frac{dy_1}{dt}\frac{dy_2}{dt}y_2K^3k_u+\frac{dy_1}{dt}\frac{dy_2}{dt}K^4k_u\\
&-\frac{dy_1}{dt}y_2^6\beta_lk_u+\frac{dy_1}{dt}y_2^5k_u(-4\beta_lK-\delta_l)+2\frac{dy_1}{dt}y_2^4Kk_u(-3\beta_lK-2\delta_l)+2\frac{dy_1}{dt}y_2^3K^2k_u(-2\beta_lK-3\delta_l)\\
&+\frac{dy_1}{dt}y_2^2K^3k_u(-\beta_lK-4\delta_l)-\frac{dy_1}{dt}y_2\delta_lK^4k_u+\frac{d^3y_2}{dt^3}y_2^5+4\frac{d^3y_2}{dt^3}y_2^4K+6\frac{d^3y_2}{dt^3}y_2^3K^2+4\frac{d^3y_2}{dt^3}y_2^2K^3\\
&+\frac{d^3y_2}{dt^3}y_2K^4-\frac{d^2y_2}{dt^2}\frac{dy_2}{dt}y_2^4-4\frac{d^2y_2}{dt^2}\frac{dy_2}{dt}y_2^3K-6\frac{d^2y_2}{dt^2}\frac{dy_2}{dt}y_2^2K^2-4\frac{d^2y_2}{dt^2}\frac{dy_2}{dt}y_2K^3-\frac{d^2y_2}{dt^2}\frac{dy_2}{dt}K^4\\
&+\frac{d^2y_2}{dt^2}y_2^6\beta_l+\frac{d^2y_2}{dt^2}y_2^5(4\beta_lK+\delta_l)+2\frac{d^2y_2}{dt^2}y_2^4K(3\beta_lK+2\delta_l)+\frac{d^2y_2}{dt^2}y_2^3K(4\beta_lK^2+c_l+6\delta_lK)\\
&+\frac{d^2y_2}{dt^2}y_2^2K^2(\beta_lK^2+2c_l+4\delta_lK)+\frac{d^2y_2}{dt^2}y_2K^3(c_l+\delta_lK)-\frac{dy_2}{dt}^2y_2^4\delta_l-4\frac{dy_2}{dt}^2y_2^3\delta_lK\\
&+3\frac{dy_2}{dt}^2y_2^2K(-c_l-2\delta_lK)-4\frac{dy_2}{dt}^2y_2K^2(c_l+\delta_lK)-\frac{dy_2}{dt}^2K^3(c_l+\delta_lK)+\frac{dy_2}{dt}y_1y_2^4\delta_lk_u\\
&+4\frac{dy_2}{dt}y_1y_2^3\delta_lKk_u+6\frac{dy_2}{dt}y_1y_2^2\delta_lK^2k_u+4\frac{dy_2}{dt}y_1y_2\delta_lK^3k_u+\frac{dy_2}{dt}y_1\delta_lK^4k_u\\
&+\frac{dy_2}{dt}y_2^6\beta_l\delta_l+4\frac{dy_2}{dt}y_2^5\beta_l\delta_lK+\frac{dy_2}{dt}y_2^4(\beta_lc_lK+6\beta_l\delta_lK^2-c_l\delta_l)+2\frac{dy_2}{dt}y_2^3K(\beta_lc_lK+2\beta_l\delta_lK^2-c_l\delta_l)\\
&+\frac{dy_2}{dt}y_2^2K^2(\beta_lc_lK+\beta_l\delta_lK^2-c_l\delta_l)-y_1y_2^6\beta_l\delta_lk_u-4y_1y_2^5\beta_l\delta_lKk_u-6y_1y_2^4\beta_l\delta_lK^2k_u\\
&-4y_1y_2^3\beta_l\delta_lK^3k_u-y_1y_2^2\beta_l\delta_lK^4k_u+y_2^6\beta_lc_l\delta_l+3y_2^5\beta_lc_l\delta_lK+3y_2^4\beta_lc_l\delta_lK^2+y_2^3\beta_lc_l\delta_lK^3= \qquad 0.
\end{aligned}
\tag{4}
$$

Equations (3) and (4) are called input–output equations of within-host model Eq. (1), which are differential polynomials involving the observed state variables $y_1 = V_u(t)$ and $y_2 = V_l(t)$ and the within-host model parameters. Note that solving input–output equations Eqs. (3) and (4) is equivalent to solving the within-host model Eq. (1) for the state variables $V_u(t)$ and $V_l(t)$. For identifiability analysis, it is crucial that the input-output equations are monic, i.e. the leading coefficient is 1. It is clear that the input–output equation Eq. (3) is monic, and the input–output equation Eq. (4) can be made monic by dividing the coefficients with the coefficient of the leading term, which is k_u. As a result, the definition of the structural identifiability within differential algebra approach which involves input–output equations takes the following form[27–30].

Definition 2 Let $c(p)$ denote the coefficients of the input-output equations, Eqs. (3) and (4), where p is the vector of model parameters. We say that the within-host model Eq. (1) is structured to reveal its parameters from the observations if and only if

$$c(p) = c(q) \quad \Longrightarrow \quad p = q.$$

Suppose $p = \{\beta_u, \delta_u, p_u, c_u, k_l, \beta_l, \delta_l, p_l, c_l, K, k_u\}$ and $q = \{\hat{\beta}_u, \hat{\delta}_u, \hat{p}_u, \hat{c}_u, \hat{k}_l, \hat{\beta}_l, \hat{\delta}_l, \hat{p}_l, \hat{c}_l, \hat{K}, \hat{k}_u\}$ are two parameter sets of the within-host model which produced the same observations. This can only happen if the coefficients of the input-output Eqs. (3) and (4) are the same. Hence, if $c(p)$ denote the coefficients of the corresponding monic polynomial of input–output equations, we solve $c(p) = c(q)$ to obtain

$$\beta_u = \hat{\beta}_u,\ \delta_u = \hat{\delta}_u,\ c_u = \hat{c}_u,\ k_u = \hat{k}_u,\ \beta_l = \hat{\beta}_l,\ \delta_l = \hat{\delta}_l,\ c_l = \hat{c}_l,\ K = \hat{K},\ k_l = \hat{k}_l. \tag{5}$$

The solution set (5) means that the parameters, $\beta_u, \delta_u, c_u, k_u, \beta_l, \delta_l, c_l, K$ and k_l can be identified uniquely. However, parameters p_u and p_l both disappear from the input–output Eqs. (3) and (4). It is easier to see the reason behind this by scaling the unobserved state variables of the within-host model Eq. (1) with a postive scalar $\sigma > 0$. Hence, $(\sigma T_u, \sigma I_u, V_u, \sigma T_l, \sigma I_l, V_l) = (\hat{T}_u, \hat{I}_u, V_u, \hat{T}_l, \hat{I}_l, V_l)$ will solve the following system

$$\begin{aligned}
\frac{d\hat{T}_u}{dt} &= -\beta_u \hat{T}_u V_u,\\
\frac{d\hat{I}_u}{dt} &= \beta_u \hat{T}_u V_u, -\delta_u \hat{I}_u,\\
\frac{dV_u}{dt} &= \hat{p}_u \hat{I}_u - c_u V_u + k_l V_l,\\
\frac{d\hat{T}_l}{dt} &= -\beta_l \hat{T}_l V_l,\\
\frac{d\hat{I}_l}{dt} &= \beta_l \hat{T}_l V_l - \delta_l \hat{I}_l,\\
\frac{dV_l}{dt} &= \hat{p}_l \hat{I}_l - c_l \frac{V_l}{V_l + K} V_l + k_u V_u,
\end{aligned} \tag{6}$$

where $\hat{p}_u = \frac{p_u}{\sigma}$ and $\hat{p}_l = \frac{p_l}{\sigma}$. Since $\sigma > 0$ was arbitrary and the observations do not give information about the scaling parameter σ, the parameters p_u and p_l can not be identified from the viral load in the URT and LRT tracts. We conclude that the within-host model Eq. (1) is not identifiable. We summarize the structural identifiability result in the following Proposition 1.

Proposition 1 *The within-host model Eq. (1) is not structured to reveal its parameters from the observations of viral load in upper and lower respiratory tracts. The parameters p_u and p_l are not identifiable and only the parameters $\beta_u, \delta_u, c_u, k_u, \beta_l, \delta_l, c_l, K, k_l$ can be identified.*

To obtain a structurally identifiable model from the V_u and V_l observations, we scale the unobserved state variables with $\hat{T}_u = p_u T_u,\ \hat{I}_u = p_u I_u,\ \hat{T}_l = p_l T_l,\ \hat{I}_l = p_l I_l$ and obtain the following scaled within-host model

$$\begin{aligned}
\frac{d\hat{T}_u}{dt} &= -\beta_u \hat{T}_u V_u,\\
\frac{d\hat{I}_u}{dt} &= \beta_u \hat{T}_u V_u, -\delta_u \hat{I}_u,\\
\frac{dV_u}{dt} &= \hat{I}_u - c_u V_u + k_l V_l,\\
\frac{d\hat{T}_l}{dt} &= -\beta_l \hat{T}_l V_l,\\
\frac{d\hat{I}_l}{dt} &= \beta_l \hat{T}_l V_l - \delta_l \hat{I}_l,\\
\frac{dV_l}{dt} &= \hat{I}_l - c_l \frac{V_l}{V_l + K} V_l + k_u V_u.
\end{aligned} \tag{7}$$

Proposition 2 *The scaled within-host model Eq. (7) is structured to reveal its parameters from the observations of viral load in upper and lower respiratory tracts. All the parameters*

$$\beta_u, \delta_u, c_u, k_u, \beta_l, \delta_l, c_l, K, k_l$$

can be identified, hence the within-host model Eq. (7) is globally identifiable.

Data fitting.

Parameter values. We assume that the upper respiratory tract susceptible population is $T_0^u = 4 \times 10^8$ epithelial cells, as in influenza studies[23]. This estimate was obtained by assuming a URT surface in adults of 160 cm^2[33] and an epithelial cell's surface area of $2 \times 10^{-11} - 4 \times 10^{-11}$ m^2[34]. We use a similar method to estimate the target cell population in the LRT. The lung's surface area of 70 m^2 (with range between 35 m^2 and 180 m^2)[35] is composed of gas exchange regions (aveoli), and of conducting airways (trachea, bronchi, bronchioles). Since the gas exchange region is affected by SARS-Cov-2 only in severe cases[20] we ignore it, and restrict the LRT compartment to the conducting airways whose surface area is 2471 ± 320 cm^2[36]. Therefore, we obtain an initial epithelial cell target population in the LRT of $T_l^0 = 6.25 \times 10^9$ epithelial cells. If we assume that viral production rates are $p_u = 50$ and $p_l = 32$ per day then, after scaling, we have initial target cell populations in the URT and LRT of $\hat{T}_u^0 = 2 \times 10^{10}$ epithelial cells and $\hat{T}_l^0 = 2 \times 10^{11}$ epithelial cells. The other initial conditions are unaffected by scaling and are set at $\hat{I}_u^0 = \hat{I}_L^0 = 0, \hat{V}_u^0 = 0.1$ and $\hat{V}_L^0 = 0$, where the virus inoculum of $\hat{V}_u^0 = 0.1$ cp/ml is set below the reported limit of quantification of 10^2 cp/ml[18]. Lastly, the incubation periods were estimated in Ref.[19] and are listed in Table 1.

Bayesian parameter estimation. During the data collection process, observations are perturbed with noise. Hence, the URT and LRT viral load deviates from the smooth trajectory of the observations $y_1(t)$ and $y_2(t)$ at measurement times. We represent measurement error using the following statistical model

$$\begin{aligned} V_u^{data}(t_i) &= y_1(t_i, \hat{p}) + \epsilon_i \quad i = 1, 2, \ldots, n_u; \\ V_l^{data}(t_j) &= y_2(t_j, \hat{p}) + \epsilon_j \quad j = 1, 2, \ldots, n_l; \end{aligned} \tag{8}$$

where $\hat{p}$ is the true parameter vector assumed to generate the data, and the random variables ϵ_i and ϵ_j are assumed to be Gaussian with mean zero and standard deviation σ.

We use Bayesian inference and Markov Chain Monte Carlo (MCMC) to determine the remaining nine parameters of the model Eq. (7)

$$p = \{\beta_u, \delta_u, c_u, k_u, \beta_l, \delta_l, c_l, K, k_l\}.$$

Bayesian inference treats model parameters as random variables and seeks to determine the parameters' posterior distribution, where the term "posterior" refers to data-informed distributions. The posterior densities are determined using Bayes' theorem, which defines them as the normalised product of the prior density and the likelihood. Let $\pi(p|\mathcal{D})$ denote the probability distribution of the parameter p given the data $\mathcal{D} = \Big(V_u(t_i), V_l(t_j)\Big)$, then the Bayes theorem states that

$$\pi(p|\mathcal{D}) = \frac{\pi(\mathcal{D}|p)\pi(p)}{\pi(\mathcal{D})},$$

where $\pi(p)$ is the prior parameter distribution and $\pi(\mathcal{D})$ is a constant which is usually considered to be a normalization constant so that the posterior distribution is indeed a probability density function (pdf), i.e. its integral equals to 1. The likelihood function $\pi(\mathcal{D}|p)$ gives the probability of observing the measurements $\mathcal{D}$ given that the parameter values is p. In terms of the within-host model Eq. (7) and the laboratory data Eq. (8), the likelihood function $\pi(\mathcal{D}|p)$ takes the following form

$$\begin{aligned} \pi(\mathcal{D}|p) = &\prod_{i=1}^{n_u} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{\sigma^2}\Big(\log_{10}(V_u(t_i)) - \log_{10}(V_u^{data}(t_i)\Big)^2} \\ &\times \prod_{j=1}^{n_l} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{\sigma^2}\Big(\log_{10}(V_l(t_j)) - \log_{10}(V_l^{data}(t_j)\Big)^2}. \end{aligned} \tag{9}$$

The ultimate goal is to determine the posterior distributions of the parameters in the light of laboratory data. To approximate the posterior distributions, we use the MCMC method introduced in Refs.[37,38]. MCMC methods generate a sequence of random samples $p_1, \ p_2, \ldots, p_N$ whose distribution asymptotically approaches the posterior distribution for size N. The random walk Metropolis algorithm is one of the most extensively used MCMC algorithms. The Metropolis algorithm starts at position p_i, then the Markov chain generates a candidate parameter value p_* from the proposal distribution, and the algorithm accepts or rejects the proposed value based on probability α given by

$$\alpha = \min\left(1, \frac{\pi(p_*)}{\pi(p_i)}\right).$$

As with the Metropolis algorithm, the essential feature of MCMC approaches is the formulation of a proposal distribution and an accept–reject criteria. In this paper, we employ the Delayed Rejection Adaptive Metropolis, (DRAM[37]) and use the MATLAB toolbox provided by the same authors[39]. In comparison to other Metropolis algorithms, the Markov chain constructed with DRAM is robust and converges rapidly (see Fig. 2).

The two patch within-host model Eq. (7) is novel, hence we do not have any prior information regarding model parameters. We determine the prior distributions by fitting the structurally identifiable within-host model Eq. (7) to patient A's data and to the population data (all nine patients). The prior distributions $\pi(p)$ are then defined as a normal distribution with a mean equal to the fitted value and variance equal to $\sigma^2, \pi(p) \sim N(\mu, \sigma)$. Table 2 shows the obtained prior distribution of each parameter for patient A and population data, together with the lower and upper bounds of the prior $\pi(p)$.

Results

Viral dynamics.

To study the kinetics of SARS-CoV-2 in the upper and lower respiratory tracts we developed a two patch within-host model Eq. (1) that assumed viral shedding between the two patches. To ensure structural identifiability, we rescaled our equations by removing the non-identifiable parameters p_u and p_l (see "Identifiability analysis" section in "Materials and methods"). The resulting model Eq. (7) was validated against SARS-CoV-2 RNA data from throat swabs and sputum samples collected from an infectious event with the same index case early in the pandemic[18]. We used Bayesian parameter estimation with the viral samples in URT and LRT from a single individual (patient A) and the entire population (nine individuals) and approximated posterior distributions with $N = 10^6$ Markov chain iterations (see "Data fitting" section in "Materials and methods").

We generated prediction graphs of the within-host model Eq. (7) by sampling parameter realizations from posterior distributions. The model's predictive posterior distribution for single patient URT-LRT viral data and

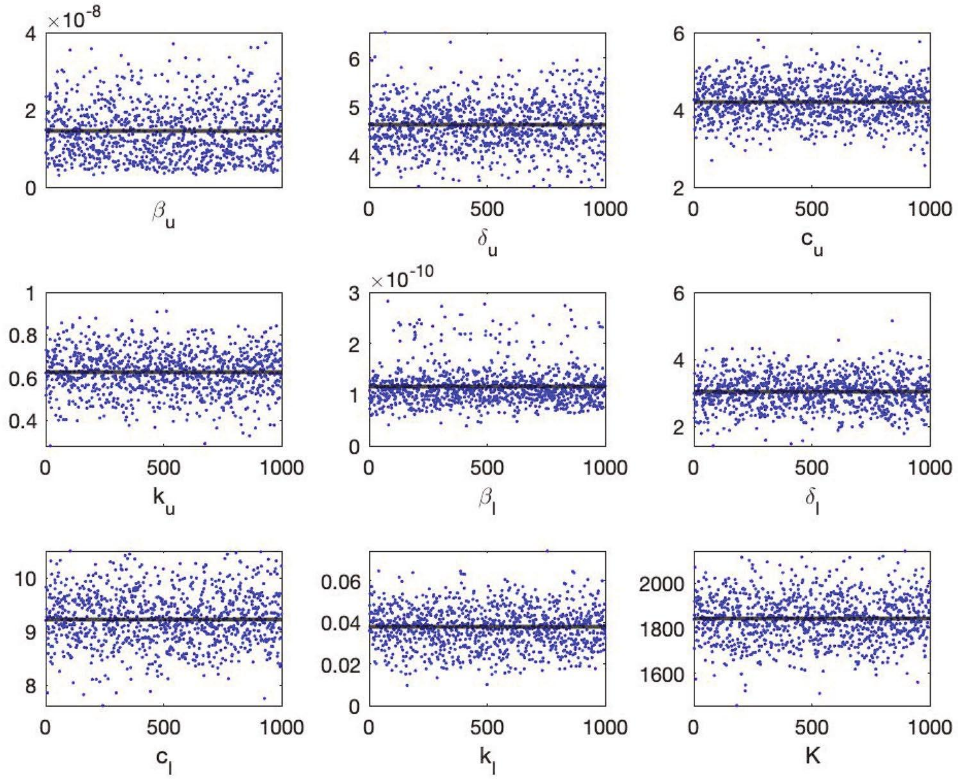

Figure 2. The Markov chain of the within-host model Eq. (7)'s parameters obtained when the model is fitted to the population data. Every 1000th point of 10^6 iterations are shown. The black line shows the mean of the chain.

		Prior $\pi(p)$	**Patient A**	**Population**
Parameter	**Description**	**Min–Max**	$\pi(p) \sim \mathcal{N}(\mu, \sigma)$	$\pi(p) \sim \mathcal{N}(\mu, \sigma)$
β_u	Viral infectivity in URT	$(10^{-12}, 10^{-7})$	$\mathcal{N}(1.1 \times 10^{-8}, 10^{-8})$	$\mathcal{N}(8.9 \times 10^{-9}, 10^{-8})$
β_l	Viral infectivity in LRT	$(10^{-12}, 10^{-7})$	$\mathcal{N}(3.9 \times 10^{-8}, 10^{-10})$	$\mathcal{N}(9.3 \times 10^{-11}, 10^{-10})$
δ_u	Infected cell decay rate in URT	$(0, 50)$	$\mathcal{N}(4.88, 0.5)$	$\mathcal{N}(4.64, 0.5)$
δ_l	Infected cell decay rate in LRT	$(0, 50)$	$\mathcal{N}(5.59, 0.5)$	$\mathcal{N}(2.99, 0.5)$
c_u	Viral decay rate in URT	$(0, 30)$	$\mathcal{N}(2.88, 0.5)$	$\mathcal{N}(4.27, 0.5)$
c_l	Viral decay rate in LRT	$(0, 30)$	$\mathcal{N}(11.43, 0.5)$	$\mathcal{N}(9.21, 0.5)$
K	V_l where loss is half-maximal	$(0, 3000)$	$\mathcal{N}(910, 100)$	$\mathcal{N}(1840, 100)$
k_u	Shedding into LRT	$(10^{-6}, 1)$	$\mathcal{N}(0.24, 0.1)$	$\mathcal{N}(0.62, 0.1)$
k_l	Shedding into URT	$(10^{-6}, 1)$	$\mathcal{N}(0.0008, 0.0001)$	$\mathcal{N}(0.036, 0.01)$

Table 2. Parameters for the within-host model Eq. (7) are listed together with their lower and upper bounds for the priors. Prior distributions are normally distributed with mean equal to the fitted value and variance, σ^2.

population URT-LRT viral data are presented in Fig. 3. The resulting dynamics show viral expansion to peak values at days 2.1 in URT and 2.9 in LRT followed by decline in both tracts (see Fig. 3). The grey areas in the graph represent the 50% and 95% posterior regions. The fewer data points in patient A results in wider model prediction range (gray regions) compared to the population predictions, especially for the LRT viral load.

While the viral titers decay to low levels (below 10^2 cp/ml) 3 weeks after infection in the URT, they stay elevated (to above 5.4×10^3 cp/ml at week four) in the LRT. To model viral RNA persistence in the LRT we included a density dependent term for the loss of LRT virus, $c_l V_l/(V_l + K)$, and estimated parameter K where V_l loss is half-maximal, together with the other viral specific terms.

We found similar mean infectivity rates in the URT for both the individual patient considered (patient A) and the entire population, $\beta_u = 1.4 \times 10^{-8}$ ml/(vir× day). By contrast, the mean infectivity rates in the LRT for patient A is 3.2-times higher than the LRT infectivity rate of the total population, $\beta_l = 3.9 \times 10^{-10}$ ml/(vir× day) versus $\beta_l = 1.2 \times 10^{-10}$ ml/(vir× day). The mean infected cells death rates are similar in URT and LRT, $\delta_u = 4.9, \delta_u = 4.6$ per day and $\delta_l = 5.7, \delta_l = 3$ per day for patient A and for the total population, respectively. The mean viral clearance rates are higher in LRT compared to URT, $c_l = 11.5, c_l = 9.2$ per day compared to $c_u = 2.8$,

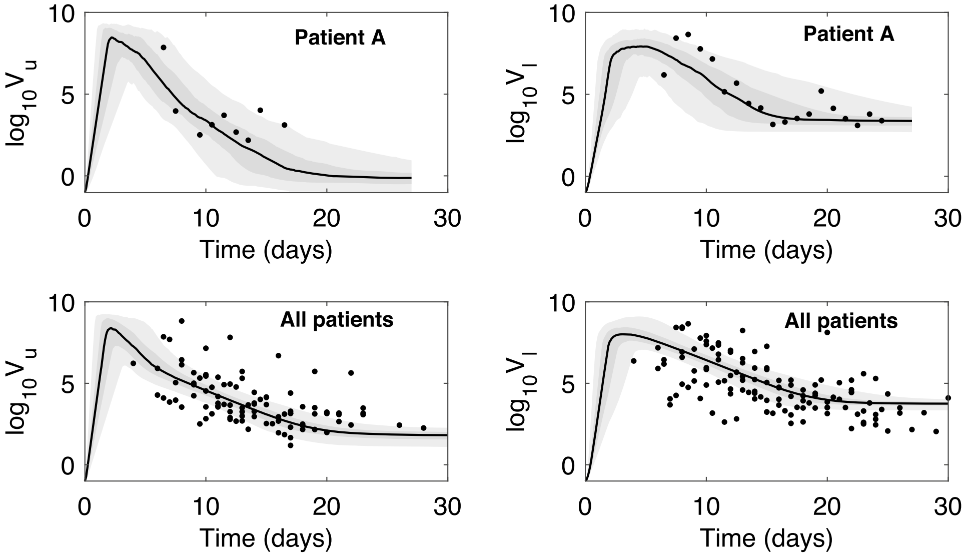

Figure 3. Virus dynamics obtained from fitting within-host model Eq. (7) to URT virus titer (left) and LRT virus titer (right) in patient A and in the entire population. The grey bars represent 50% and 95% posterior regions.

$c_u = 4.2$ per day, for patient A and for the total population, respectively. This may indicate increased immune responses occurring in LRT. The mean URT to LRT shedding rates are higher than the mean LRT to URT shedding rates, $k_u = 0.24$, $k_u = 0.63$ (swab/ml) per day compared to $k_l = 7.9 \times 10^{-4}$, $k_l = 0.04$ (ml/swab) per day for patient A and for the total population, respectively. This one way shedding was observed by other studies that investigated the Wofle et al. data[3]. Lastly, the mean LRT viral load where viral clearance is half-maximal is $K = 910$ RNA per ml for patient A and $K = 1841$ RNA per ml for the total population.

Practical identifiability. During MCMC data fitting, we used parameters limits predetermined to range around a single point estimation obtained using the 'fminsearch' algorithm in Matlab (see Table 2). Parameter distributions for the nine parameter considered $p = \{\beta_u, \delta_u, c_u, k_u, \beta_l, \delta_l, c_l, K, k_l\}$ were obtained using an MCMC Bayesian approach that sampled the parameter space $N = 10^6$ times. We apply DRAM MCMC algorithm and observe fast convergence of the chains (see Fig. 2). The resulting distributions, together with the prior probability density functions (pdf) are presented in Fig. 4. We observe good agreement between the prior pdf and the posterior distributions for all parameters with the exception of infectivity rates β_u and β_l. Moreover, while all parameters follow normal distributions for patient A (Fig. 4A and Supplementary Fig. S1), the LRT infectivity rate β_l follows a bimodal distribution in the fit to the total population data (Fig. 4B and Supplementary Fig. S2).

Figure 5 shows the scatter plots of for paired (β_l, k_l), (β_l, K), (β_l, δ_l) and (β_l, c_l) parameter distributions obtained when the within-host model Eq. (7) is fitted to patient A's data (panel A) and population data (panel B) (see also Supplementary Fig. S2 for the scatter plots of all parameter distributions). In the scatter plots for the population data containing parameter β_l we observe bimodal clustering. In joint density plots, bimodal clustering may suggest practical unidentifiability[40]. This implies that, despite the fact that we have shown that the within-host model Eq. (7) is structurally identifiable, it may in fact not be practically unidentifiable. It is well understood that a structurally identifiable model may be practically unidentifiable[28–30,41]. Many variables can lead to practical non-identifiability, such as considerable noise in the data, limited data points, or inadequate timing of data collection.

Optimal experimental design. The possible lack of practical identifiability for the total population may be due to (1) restrictions on the parameter space and the types of distributions we are imposing on the parameters, or (2) the limited data points early in the infection.

To investigate the first hypothesis, we collected samples in the parameter space of

$$\{\ln \beta_u, \ln \beta_l, \ln k_u, \ln k_l\},$$

rather than $\{\beta_u, \beta_l, k_u, k_l\}$ and the assumed that either $\{\ln \beta_u, \ln \beta_l, \ln k_u, \ln k_l\}$ are normally distributed, or that $\{\beta_u, \beta_l, k_u, k_l\}$ are lognormally distributed. We set the limits of logarithmic parameter priors as in Table 3 while keeping the limits of the other parameters as before (see Table 2). We sampled the new parameter space $N = 10^6$ times and reapplied the MCMC Bayesian approach. The resulting estimates for parameters $p = \{\beta_u, \delta_u, c_u, k_u, \beta_l, \delta_l, c_l, K, k_l\}$ no longer show bimodal results regardless on whether we assume that $\{\ln \beta_u, \ln \beta_l, \ln k_u, \ln k_l\}$ are normally distributed (see Fig. 6A) or that $\{\beta_u, \beta_l, k_u, k_l\}$ are lognormally distributed (see Fig. 6B).

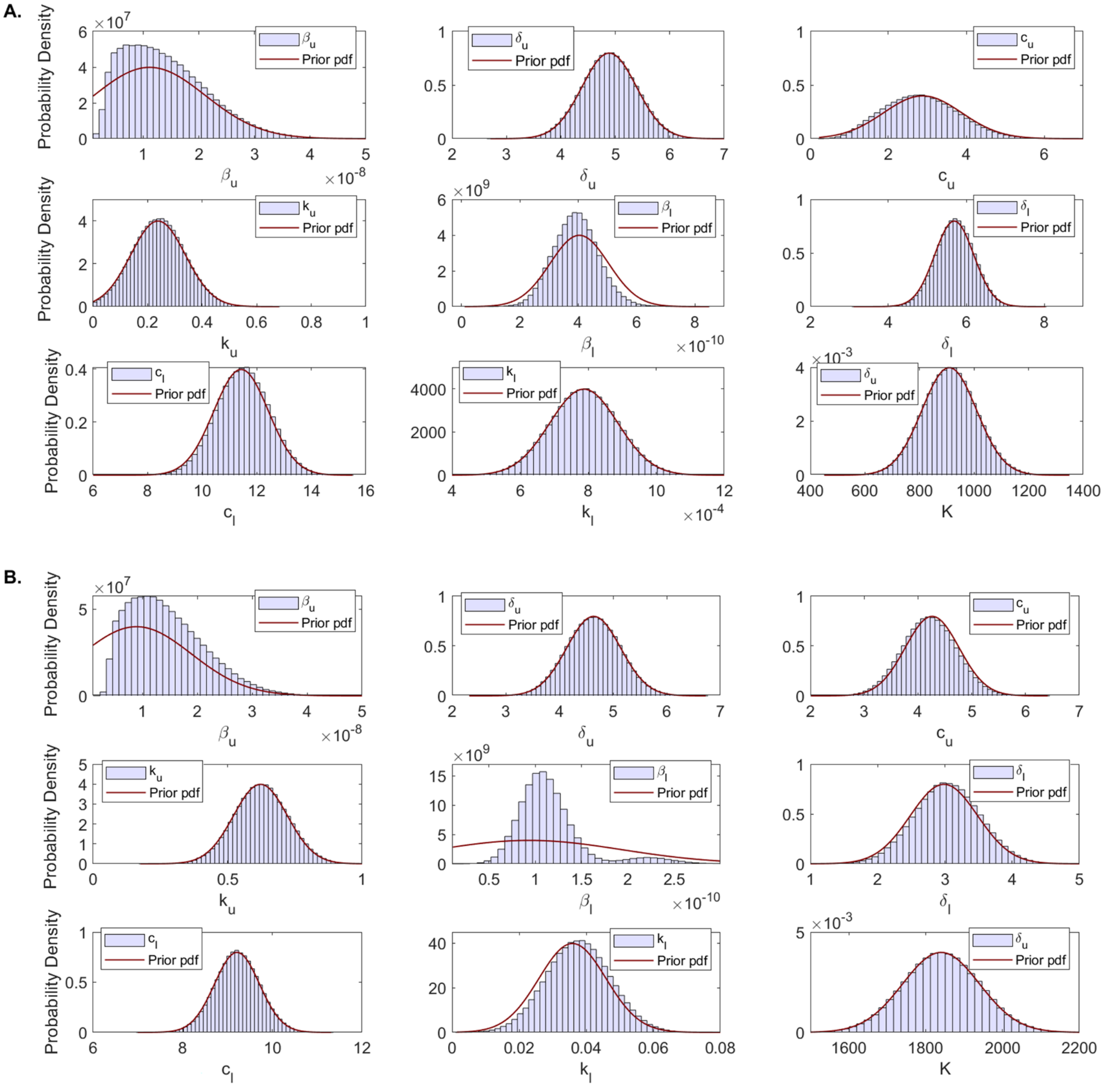

Figure 4. Histogram of estimated parameter distributions from fitting within-host model Eq. (7) to URT virus titer and LRT virus titer in: (**A**) patient A and (**B**) entire population. All parameters were considered normally distributed.

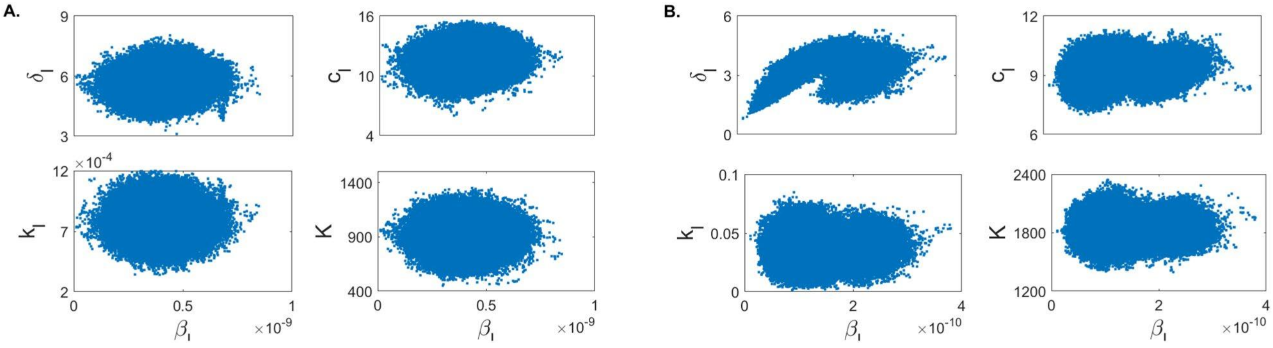

Figure 5. Scatter plots showing correlation among relevant parameters for (**A**) patient A and (**B**) total population.

Parameter	Description	Prior $\pi(p)$	Population
		Min–Max	$\pi(p) \sim \mathcal{N}(\mu, \sigma)$
$\ln \beta_u$	Viral infectivity in URT (log scale)	$(-25, -14)$	$\mathcal{N}(-18.5, 1)$
$\ln \beta_l$	Viral infectivity in LRT (log scale)	$(-30, -15)$	$\mathcal{N}(-23, 1)$
$\ln k_u$	Shedding into LRT (log scale)	$(-11, 0)$	$\mathcal{N}(-0.5, 1)$
$\ln k_l$	Shedding into URT (log scale)	$(-7, 4)$	$\mathcal{N}(-3.3, 1)$

Table 3. Adjusted parameters for the within-host model Eq. (7) are listed together with their lower and upper bounds for the priors which are normally distributed with mean equal to the fitted value and variance, σ^2.

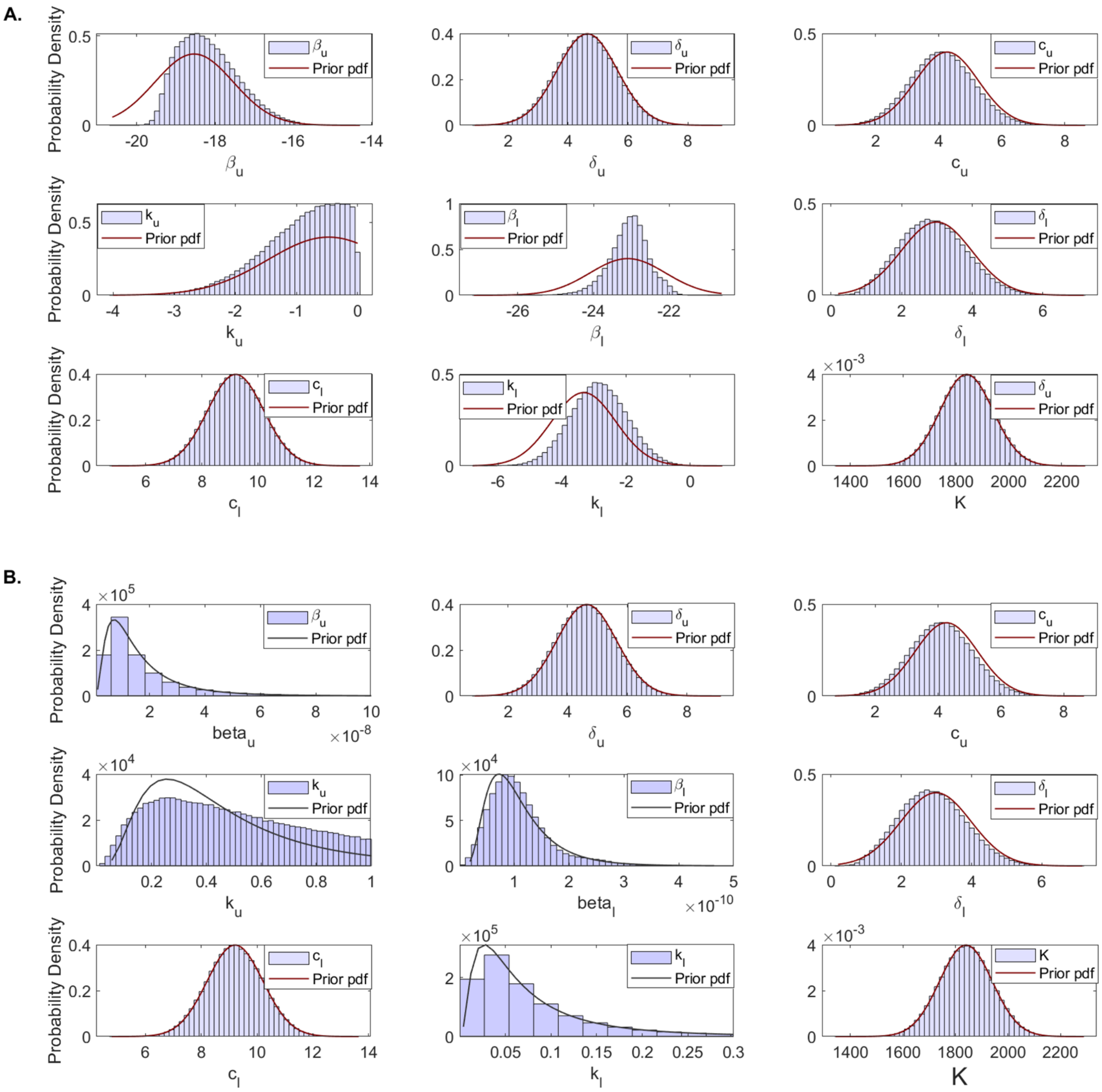

Figure 6. Histogram of estimated parameter distributions from fitting model Eq. (7) to URT virus titer and LRT virus titer in total populations. (**A**) Parameters $\ln \beta_u$, $\ln \beta_l$, $\ln k_u$, $\ln k_l$ were considered normally distributed. (**B**) Parameters β_u, β_l, k_u, k_l were considered lognormally distributed. All other parameters were considered normally distributed.

To investigate the second hypothesis, we created synthetic data and used it to further examine how the timing of the data collection in the population correlates to the structure of the resulting parameter estimations. We assumed that the real data corresponds to the solution of model Eq. (7) with parameters in Tables 2 and 3 which are randomly perturbed according to Eq. (8) with errors ϵ_i and ϵ_j assumed to be uniformly distributed with mean 0 and standard deviation 0.5. We produced two data sets. The first data set, which assumes data is collected daily from day 0 to day 12 post infection, is

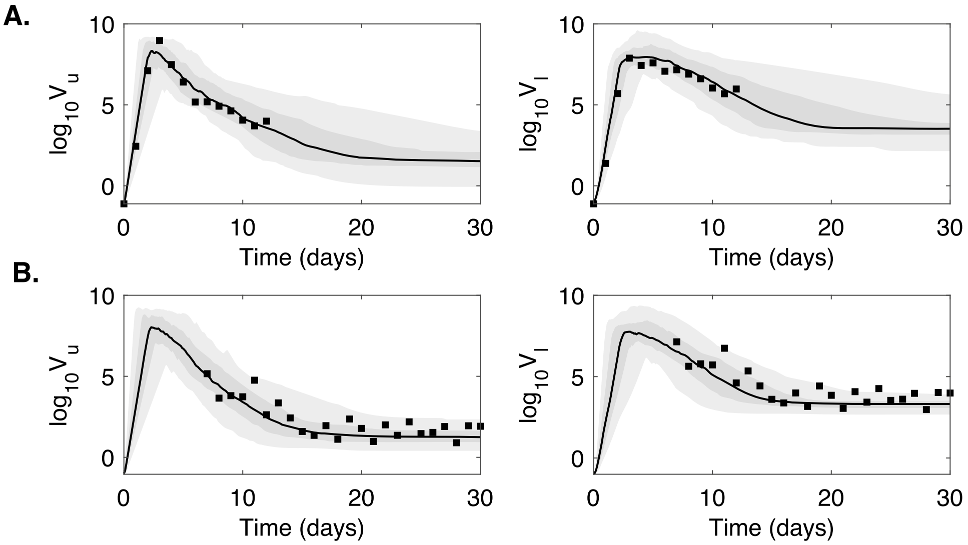

Figure 7. Virus dynamics obtained from fitting within-host model Eq. (7) to URT virus titer and LRT virus titer in (**A**) Experiment 1 and (**B**) Experiment 2. The grey bars represent 50% and 95% posterior regions.

$$\textbf{Experiment1}: \left(V_u^{data}(t_j), V_l^{data}(t_j)\right) \text{ for } t_j = \{1, \ldots, 12\}.$$

The second data set, which assumes data is collected from day 7 to day 27 post infection, is

$$\textbf{Experiment2}: \left(V_u^{data}(t_j), V_l^{data}(t_j)\right) \text{ for } t_j = \{7, \ldots, 27\}.$$

Since the practical identifiability is a local property of the parameters, we used the priors for

$$p = \{\ln \beta_u, \ln \beta_l, \ln k_u, \ln k_l\}$$

given in Table 3 and the priors for the rest of the parameters as in Table 2, to generate prediction graphs of the within-host model Eq. (7). The model's predictive posterior distribution for all patients' URT- LRT viral data for Experiments 1 and 2 are presented in Fig. 7 together with grey areas for the 50% and 95% posterior regions (see also Supplementary Fig. S4). As expected, we observe wider model prediction ranges (gray regions) in the second phase decay for experiment 1 and in the expansion and peak areas for experiment 2, where data is scarce (Fig. 7).

To determine whether practical identifiability is lost in each experiment we created parameter histograms for each parameters (see Fig. 8 and Supplementary Fig. S3). When data samples at the expansion stages of the infection are collected (as in Experiment 1), the LRT infectivity parameter β_l follows a normal distribution (see Fig. 8A, left panel, blue bars). This results are validated by the corresponding dual parameter scatter plots (see Fig. 9A). In contrast, when the data at the expansion stages of the infection is sparse (as in Experiment 2), the LRT infectivity parameter β_l follows a bimodal distribution (see Fig. 8A, right panel, blue bars). This results are observed in the corresponding scatter plots, where we see bimodal clustering involving not just parameter β_l, but involving parameter β_u as well (see Fig. 9B). These results can be slightly improved when we consider that β_u and β_l follow lognormal distributions (see Fig. 8B, right panel, blue bars). This suggests that the practical unidentifiability that appeared in the population data might be fixed by collecting data at the early stages of infection.

Discussion

In this study, we developed a within-host mathematical model of SARS-CoV-2 infection that connected the virus kinetics in the upper and lower respiratory tracts of infected individuals and used it to determine the tract specific viral parameters. We removed viral production rates, to ensure structural identifiability, and fitted the rescaled model Eq. (7) to published longitudinal throat swabs and sputum titers in a single individual and in the entire population from SARS-CoV-2 infection study[18]. We estimated nine unknown parameters using an MCMC Bayesian fitting approach[37]. To avoid over fitting, we determined best estimates in a single patient (for which we have 26 data points) and in the entire population (for which we have 201 data points). We found shorter virus life-spans in LRT compared to viral URT, 2–3 h compared to 5.7–8.5 h. Our LRT estimates are similar to the fixed (and non-tract specific) virus life-span of 2.4 h used by Ke et al.[3] and the estimated (and tract specific) life-span of 1.2 h in Wang et al.[6], but longer than the 10 h seen in influenza and used by Hernandez et al.[7]. The between tracts differences may suggest the presence of additional immune mediated viral clearance in the LRT. We found similar infected cells life-span between the two tracts, with a range of 4.2–8 h, shorter than in other studies[3,7]. Lastly, the mean URT basic reproductive number for the entire population, $R_0 = \frac{\beta T_0}{c\delta}$, equals 17.4, higher than in Ref.[3]. While we assumed two-way viral shedding between tracts, data fitting suggested higher virus shedding from upper and lower respiratory tracts than the other way around, consistent with other studies[3].

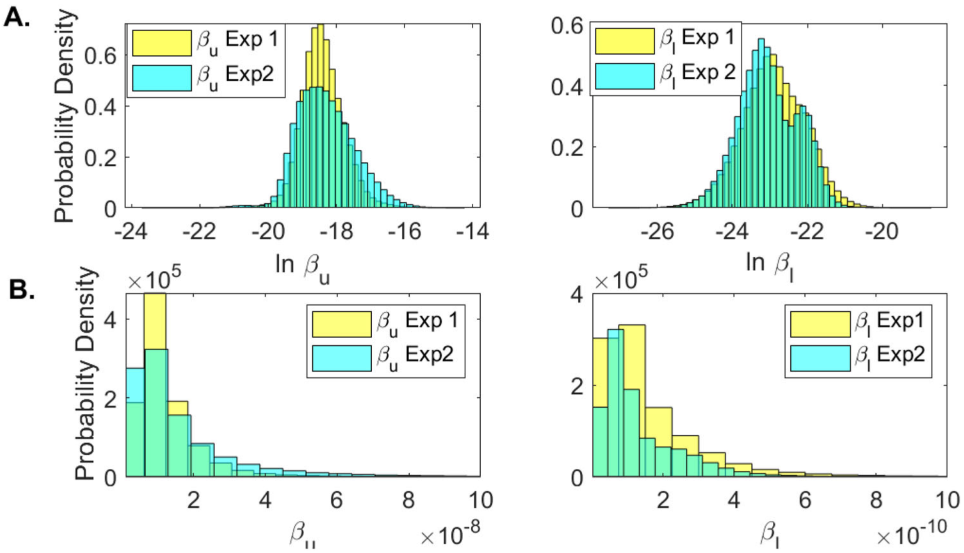

Figure 8. Histograms for β_u and β_l for Experiment 1 (yellow) and Experiment 2 (blue) when (**A**) ln β_u and ln β_l are assumed to be normally distributed; and (**B**) β_u and β_l are assumed to be lognormally distributed.

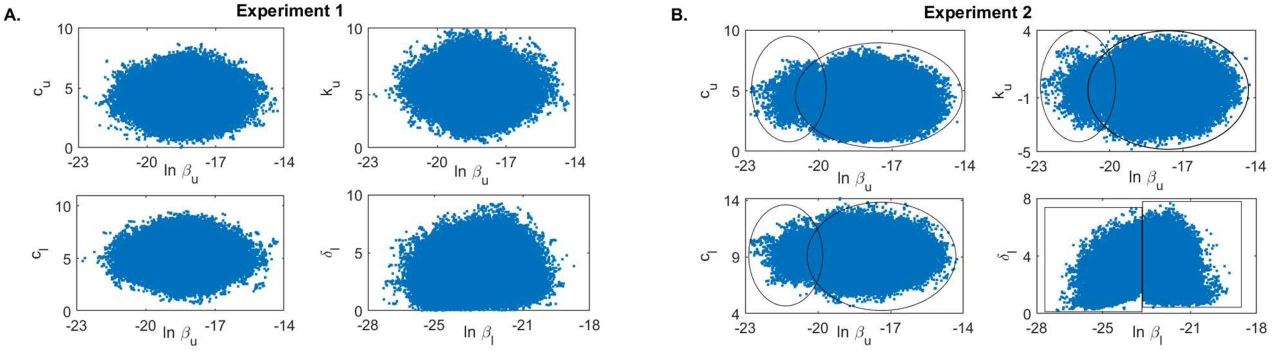

Figure 9. Scatter plots for (**A**) Experiment 1 and (**B**) Experiment 2. Parameters ln β_u and ln β_l are assumed to be normally distributed.

Interestingly, we found that the estimated LRT infectivity rate parameter follows a bimodal distribution when the model was fitted to the entire population data. We attributed this behavior to practical non-identifiability. Practical non-identifiability is observed when the measured data is contaminated with noise. We have inherently accounted for noisy data by combining viral measurements from nine patients with different viral profiles. We investigated several ways for improving practical identifiability of this parameter and found that both estimating the logaritmic value ln β_l and assuming log-normally distributions for some parameters improves the accuracy of our estimates.

Most importantly, it has been reported that practical identifiability can be achieved by adding pertinent data measurements that can help improve the identity of unknown parameters[42,43]. Such a process, known as optimal experimental design, aims to obtain additional information about a system through the addition of new measurements. Since in system Eq. (7) the non-practically identifiable infectivity parameter β_l is responsible for the LRT dynamics early in the infection, we investigated whether the addition of early data contains the maximal information needed for improving its estimate. We created two virtual data sets, one in which data is collected daily for the first 12 days and one in which data is collected daily for 20 days, starting at day 7. We found that the infectivity rate β_l is bimodal and, hence, non-practically identifiable when data is missing during the first 7 days of infection. The absence of early data leads to an underestimation of overall LRT viral titer in the first 14 days following infection (see Supplementary Fig. S4). This may affect one's ability for determining the best window for antiviral and immune modulation interventions[44]. Moreover, it will provide an underestimate for the period of maximum infectiousness[14], which may affect recommendations for quarantine and isolation[1]. Hence, the existence of data measurements before and/or at symptoms onset is crucial for best parameter estimation and model prediction when considering noisy population data.

Our study has several limitations. First, we considered a density dependent clearance term for the URT virus that saturates at around 1–2 × 10^3 RNA copies per ml, in order to explain the viral RNA persistence in the LRT at 30 days following infection reported in the Wolfle et al.[18]. While in public health setting a SARS-CoV-2 diagnostic is determined by PCR assays, long-term RNA levels are not a reliable measurement of infectiousness, with the measured RNA values indicating the presence of genomic fragments, immune-complexed or neutralised virus, rather than replication-competent virus[14,45,46]. Further work is needed to separate the presence of infectious versus non-infectious viral RNA in the lower respiratory tract. Secondly, we did not consider an eclipse phase in the virus infectiousness (usually assumed to be around 6 h[3,14]). This simplification may be the leading reason for larger infected cell death rate estimates in our study compared to other studies[3,7]. Thirdly, due to the novelty of the model, we have no information on parameter priors. Therefore, we fitted the within-host model to the patient A and population data, and used those estimates as means in the prior distributions. However, since the resulting means fall within ranges observed for other acute infections[23–26,47,48], and since we consider large standard deviations around the prior means, we are confident that we are covering a large enough search space that does not exclude viable outcomes.

In conclusion, we have developed a within-host model of SARS-CoV-2 infection in the upper and lower respiratory tracts, used it to determine pertinent viral parameters, and suggested the optimal experimental designs that can help improve the model predictions. These techniques may inform interventions.

Data availibility

The code generated during the current study will be made available on the corresponding author's page upon acceptance.

Received: 8 May 2022; Accepted: 16 August 2022
Published online: 27 August 2022

References

1. Forde, J. E. & Ciupe, S. M. Quantification of the tradeoff between test sensitivity and test frequency in a COVID-19 epidemic-a multi-scale modeling approach. *Viruses* **13**, 457 (2021).
2. Forde, J. E. & Ciupe, S. M. Modeling the influence of vaccine administration on COVID-19 testing strategies. *Viruses* **13**, 2546 (2021).
3. Ke, R., Zitzmann, C., Ho, D. D., Ribeiro, R. M. & Perelson, A. S. In vivo kinetics of SARS-CoV-2 infection and its relationship with a person's infectiousness. *Proc. Natl. Acad. Sci.* **118**, e2111477118 (2021).
4. Kim, K. S. *et al.* A quantitative model used to compare within-host SARS-CoV-2, MERS-CoV, and SARS-CoV dynamics provides insights into the pathogenesis and treatment of SARS-CoV-2. *PLoS Biol.* **19**, e3001128 (2021).
5. Sadria, M. & Layton, A. T. Modeling within-host SARS-CoV-2 infection dynamics and potential treatments. *Viruses* **13**, 1141 (2021).
6. Wang, S. *et al.* Modeling the viral dynamics of SARS-CoV-2 infection. *Math. Biosci.* **328**, 108438 (2020).
7. Hernandez-Vargas, E. A. & Velasco-Hernandez, J. X. In-host mathematical modelling of COVID-19 in humans. *Annu. Rev. Control.* **50**, 448–456 (2020).
8. Jenner, A. L. *et al.* COVID-19 virtual patient cohort suggests immune mechanisms driving disease outcomes. *PLoS Pathog.* **17**, e1009753 (2021).
9. Gonçalves, A. *et al.* SARS-CoV-2 viral dynamics in non-human primates. *PLoS Comput. Biol.* **17**, e1008785 (2021).
10. Vaidya, N. K., Bloomquist, A. & Perelson, A. S. Modeling within-host dynamics of SARS-CoV-2 infection: A case study in ferrets. *Viruses* **13**, 1635 (2021).
11. Goyal, A., Reeves, D. B., Cardozo-Ojeda, E. F., Schiffer, J. T. & Mayer, B. T. Viral load and contact heterogeneity predict SARS-CoV-2 transmission and super-spreading events. *Elife* **10**, e63537 (2021).
12. Ke, R. *et al.* Daily longitudinal sampling of SARS-CoV-2 infection reveals substantial heterogeneity in infectiousness. *Nat. Microbiol.* **7**, 640–652 (2022).
13. Buetti, N. *et al.* SARS-CoV-2 detection in the lower respiratory tract of invasively ventilated ARDS patients. *Crit. Care* **24**, 1–6 (2020).
14. Heitzman-Breen, N. & Ciupe, S. Modeling within-host and aerosol dynamics of SARS-CoV-2: The relationship with infectiousness. *bioRxiv* (2022).
15. Néant, N. *et al.* Modeling SARS-CoV-2 viral kinetics and association with mortality in hospitalized patients from the French COVID cohort. *Proc. Natl. Acad. Sci.* **118**, e2017962118 (2021).
16. Padmanabhan, P., Desikan, R. & Dixit, N. M. Modeling how antibody responses may determine the efficacy of covid-19 vaccines. *Nat. Comput. Sci.* **2**, 123–131 (2022).
17. Goyal, A. *et al.* Slight reduction in SARS-CoV-2 exposure viral load due to masking results in a significant reduction in transmission with widespread implementation. *Sci. Rep.* **11**, 1–12 (2021).
18. Wölfel, R. *et al.* Virological assessment of hospitalized patients with covid-2019. *Nature* **581**, 465–469 (2020).
19. Böhmer, M. M. *et al.* Investigation of a covid-19 outbreak in Germany resulting from a single travel-associated primary case: A case series. *Lancet. Infect. Dis.* **20**, 920–928 (2020).
20. Mason, R. J. Pathogenesis of covid-19 from a cell biology perspective. *Eur. Respir. J.* **55**, 1–3 (2020).
21. Pan, Y., Zhang, D., Yang, P., Poon, L. L. & Wang, Q. Viral load of SARS-CoV-2 in clinical samples. *Lancet. Infect. Dis.* **20**, 411–412 (2020).
22. To, K.K.-W. *et al.* Temporal profiles of viral load in posterior oropharyngeal saliva samples and serum antibody responses during infection by SARS-CoV-2: An observational cohort study. *Lancet. Infect. Dis.* **20**, 565–574 (2020).
23. Baccam, P., Beauchemin, C., Macken, C. A., Hayden, F. G. & Perelson, A. S. Kinetics of influenza A virus infection in humans. *J. Virol.* **80**, 7590–7599 (2006).
24. Beauchemin, C. A. & Handel, A. A review of mathematical models of influenza A infections within a host or cell culture: Lessons learned and challenges ahead. *BMC Public Health* **11**, S7 (2011).
25. Best, K. *et al.* Zika plasma viral dynamics in nonhuman primates provides insights into early infection and antiviral strategies. *Proc. Natl. Acad. Sci.* **114**, 8847–8852 (2017).
26. Nikin-Beers, R. & Ciupe, S. M. The role of antibody in enhancing dengue virus infection. *Math. Biosci.* **263**, 83–92. https://doi.org/10.1016/j.mbs.2015.02.004 (2015).
27. Miao, H., Xia, X., Perelson, A. S. & Wu, H. On the identifiability of nonlinear ode models and applications in viral dynamics. *SIAM Rev.* **53**, 3–39 (2011).
28. Tuncer, N. & Le, T. Structural and practical identifiability analysis of outbreak models. *Math. Biosci.* **299**, 1–18 (2018).
29. Tuncer, N. & Martcheva, M. Determining reliable parameter estimates for within-host and within-vector models of Zika virus. *J. Biol. Dyn.* **15**, 430–454 (2021).

30. Eisenberg, M. C., Robertson, S. L. & Tien, J. H. Identifiability and estimation of multiple transmission pathways in cholera and waterborne disease. *J. Theor. Biol.* **324**, 84–102 (2013).
31. Stigter, J., Joubert, D. & Molenaar, J. Observability of complex systems: Finding the gap. *Sci. Rep.* **7**, 1–9 (2017).
32. Bellu, G., Saccomani, M. P., Audoly, S. & D'Angiò, L. Daisy: A new software tool to test global identifiability of biological and physiological systems. *Comput. Methods Progr. Biomed.* **88**, 52–61 (2007).
33. Menache, M. *et al.* Upper respiratory tract surface areas and volumes of laboratory animals and humans: Considerations for dosimetry models. *J. Toxicol. Environ. Health* **50**, 475–506 (1997).
34. Fedoseev, G. & Geharev, S. Basic defense mechanisms of bronchio-lung system. *Gen. Pulmonol.* **1**, 112–144 (1989).
35. Fröhlich, E., Mercuri, A., Wu, S. & Salar-Behzadi, S. Measurements of deposition, lung surface area and lung fluid for simulation of inhaled compounds. *Front. Pharmacol.* **7**, 181 (2016).
36. Mercer, R. R., Russell, M. L., Roggli, V. L. & Crapo, J. D. Cell number and distribution in human and rat airways. *Am. J. Respir. Cell Mol. Biol.* **10**, 613–624 (1994).
37. Haario, H., Laine, M. & Mira, A. D. R. A. M. Efficient adaptive MCMC. *Stat. Comput.* **16**, 339–354 (2006).
38. Haario, H., Saksman, E. & Tamminen, J. An adaptive metropolis algorithm. *Bernoulli* **7**, 223–242 (2001).
39. Laine, M. Mcmc toolbox for Matlab. https://mjlaine.github.io/mcmcstat/. (accessed 18 Feb 2022).
40. Siekmann, I., Sneyd, J. & Crampin, E. J. MCMC can detect nonidentifiable models. *Biophys. J.* **103**, 2275–2286 (2012).
41. Tuncer, N., Martcheva, M., Labarre, B. & Payoute, S. Structural and practical identifiability analysis of Zika epidemiological models. *Bull. Math. Biol.* **80**, 2209–2241 (2018).
42. Wieland, F.-G., Hauber, A. L., Rosenblatt, M., Tönsing, C. & Timmer, J. On structural and practical identifiability. *Curr. Opin. Syst. Biol.* **25**, 60–69 (2021).
43. Franceschini, G. & Macchietto, S. Model-based design of experiments for parameter precision: State of the art. *Chem. Eng. Sci.* **63**, 4846–4872 (2008).
44. Caraco, Y. *et al.* Phase 2/3 trial of molnupiravir for treatment of Covid-19 in nonhospitalized adults. *NEJM Evid.* **1**, EVIDoa2100043 (2022).
45. Andersson, M. I. *et al.* SARS-CoV-2 RNA detected in blood products from patients with COVID-19 is not associated with infectious virus. *Wellcome Open Res.* **5** (2020).
46. Puhach, O. *et al.* Infectious viral load in unvaccinated and vaccinated individuals infected with ancestral, Delta or Omicron SARS-CoV-2. *Nat. Med.* **28**, 1491–1500 (2022).
47. Nikin-Beers, R. & Ciupe, S. M. Modelling original antigenic sin in dengue viral infection. *Math. Med. Biol. J. IMA* **35**, 257–272. https://doi.org/10.1093/imammb/dqx002 (2017).
48. Banerjee, S., Guedj, J., Ribeiro, R. M., Moses, M. & Perelson, A. S. Estimating biologically relevant parameters under uncertainty for experimental within-host murine West Nile virus infection. *J. R. Soc. Interface* **13**, 20160130 (2016).

Acknowledgements

SMC acknowledges support from National Science Foundation Grants No. DMS-1813011 and DMS-2051820 and by the Virginia Tech Center for Emerging, Zoonotic, and Arthropod-borne Pathogens (CeZAP) seed Grant. NT acknowledges partial support from National Science Foundation Grant DMS-1951626.

Author contributions

All authors conceived the study and performed the analyses. N.T. wrote the code. S.M.C. wrote the manuscript. All authors reviewed and revised the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-18683-x.

Correspondence and requests for materials should be addressed to S.M.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.