Journal of Applied Analysis and Computation
Volume 12, Number 4, August 2022, 1475–1495
Website:http://www.jaac-online.com
DOI:10.11948/20210300

STRUCTURAL AND PRACTICAL IDENTIFIABILITY ANALYSES ON THE TRANSMISSION DYNAMICS OF COVID-19 IN THE UNITED STATES*

Hui Wu[1], Yafei Zhao[1], Chen Zhang[3], Jianhong Wu[2,†] and Jie Lou[1,†]

Abstract We formulate an epidemic model to capture essential epidemiology of COVID-19 and major public health interventions. We start with a system of differential equations involving six compartments, and we use the Goodman and Weare affine invariant ensemble Markov Chain Monte Carlo algorithm (GWMCMC) to identify a simplified version of the full model that consists of only four compartments. We examine well-posedness of the relevant parameter estimation problem for the given observations using the U.S. epidemic data; study the reliability of model selection; analyze the structural identifiability of the selected model; and conduct a practical identifiability analysis on the selected model using the GWMCMC algorithm. Our study shows that the selected model is structurally identifiable for the confirmed cases, and for small measurement errors, key parameters such as the transmission rate are practically identifiable. We also analyze the stability of the selected model and prove the global asymptotic stability of the disease-free equilibrium and the endemic equilibrium by constructing appropriate Lyapunov functions. Our numerical experiments show that the U.S. will undergo damped transit oscillations towards the endemicity.

Keywords Dynamic model, COVID-19, model selection, identifiability analysis

MSC(2010) 92D30, 92D40.

1. Introduction

COVID-19 emerged in late 2019 has caused a pandemic, and triggered multiple waves of mathematical modeling activities. Differential equation models have been used to estimate the key ecological parameters, to understand the epidemiological process, and to evaluate effectiveness of control measures [3,4,11,16,18,19,27,28,35].

†The corresponding author.
Email: jie.lou@126.com(J. Lou), wujh@yorku.ca(J. Wu)
[1]Department of Mathematics, Shanghai University, 99 Shangda Road Shanghai, 200444, China
[2]Laboratory for Industrial and Applied Mathematics, York University, Toronto M3J 1P3, Canada
[3]School of Nursing, University of Rochester, Rochester, NY, USA
*The authors were supported by National Natural Science Foundation of China (No. 11771277).

A standard method of using dynamical models to study epidemic trends is to formulate a model mechanistically based upon the epidemiological characteristics and the available data before calibrating the model. This method is criticized for its ignoring data characteristics leading to concern on credibility of analytical results from the model. Model selection criteria have been developed to strengthen the model's credibility and generalizability. The Akaike information criterion [1] and The Bayesian information criterion [25] are two common criteria for model selection. For example, Eisenberg et al. [8] applied the SIWR model to data obtained from a cholera outbreak in 2006 in Angola. They used model selection to compare the relative importance of the direct or indirect environmental/waterborne transmission routes in the outbreak. The results showed that both direct and indirect modes of transmission were important for explaining the Angola outbreak. Model selection is often used in microscopic models to test the biological assumption to determine if it fits the experimental data [21, 33]. Therefore, an optimal model that can best represent observed data and can be used for sequent analyses and predictions can provide credible information to inform decision making.

Parameter values directly impact on the reliability of an established mathematical model. Values of key parameters are obtained either through the laboratory or data fitting. Due to the limited availability of surveillance data, we can only fit the finite data to estimate parameters in the model. Therefore, it may lead to the non-uniqueness of the parameter values to cause parameter non-identifiability (i.e., there are a variety of parameter combinations that will achieve the best fitting effect between the model and data, the prediction results produced by different parameter combinations may be very different). Therefore, it is very important to explore the well-posedness of the parameter estimation problem for any given observations in an epidemic model. If a model is not identifiable, it is difficult to guarantee the reliability of the estimated parameters.

Model parameters' identifiability analyses include structural and practical identifiability. The former is carried out under the assumption that the data are noise-free [2, 5, 8, 22]. The latter, on the other hand, is carried out when the data is affected by noise [8, 20, 30]. For any given large set of error-free data points, the model is considered structurally identifiable if it is theoretically possible to uniquely determine the parameter values from these observations. However, the model is not structurally identifiable if two or more parameter sets can lead to the same observational output. On the other hand, a structurally identifiable model may not be practically identifiable. It may be defined as the ability to estimate a given set of parameters with an accuracy considered satisfactory according to the context of the study [13, 29]. In a practical identifiability analysis, if the estimated values of the parameter are not sensitive to measurement errors and can always be well estimated, we say that the estimated value of the parameter is "acceptable" or practically identifiable. If the estimated values of the parameter are quite sensitive to measurement errors, we claim that they are "unacceptable" or practically unidentifiable [22]. For example, Tuncer et al. [30] performed both structural and practical identifiability analyses to classical epidemic models such as SIR, SEIR, and epidemic model with the treatment class (SITR), using different types of data sets. These findings suggest that health agencies should report prevalence rather than incidence for the best result of model identifiability.

In general, there are two optimization schemes for non-identifiable parameters: The first approach is to add independent data sets and use multiple data to fit the

parameters [31], and another approach is to adopt better fitting algorithms [24]. The Goodman and Weare affine invariant ensemble Markov Chain Monte Carlo algorithm(GWMCMC) is the most efficient fitting method nowadays [12,23,34]. The algorithm combines affine invariant ensemble with the traditional MCMC algorithm. It uses many walkers to iterate simultaneously, and the next position for any one walker is suggested by an affine invariant transformation which is constructed using the current positions of the other walkers [12]. Its performance is not affected by affine transformations of space. Even in the face of non-identifiability, it can converge to the target posterior distribution quickly.

In this paper, we formulate a general ODE compartmental model based on COVID-19 epidemiology and data from the United States(U.S.) to carry out model selection and identifiability analysis to predict the epidemic in the U.S. In the following section, a general ODE compartmental model (including four special cases) is formulated, and the AICc criterion is used for model selection. In Section 3, we analyze the structural and practical identifiability of the best model. In Section 4, considering the changes of control measures, we add the social distance term and vaccination to the model. Some unknown parameters of the model have fitted again, and compare the epidemic trends in the United States under different scenarios.

2. The transmission dynamics with constant recruitment: global convergence to endemic states

In this section, we build a general ODE compartmental model for COVID-19. Let $S(t)$, $E(t)$, $A(t)$, $I(t)$, $Q(t)$ and $R(t)$ denote the number of susceptible, exposed, asymptomatic, symptomatic individuals who are not diagnosed, diagnosed and recovered individuals at time t respectively. In this model, we consider the population's constant recruitment rate and allow recovery individuals to be susceptible since some recovered individuals have a low level of antibody [14]. The compartment diagram of the model is shown in Fig 1.

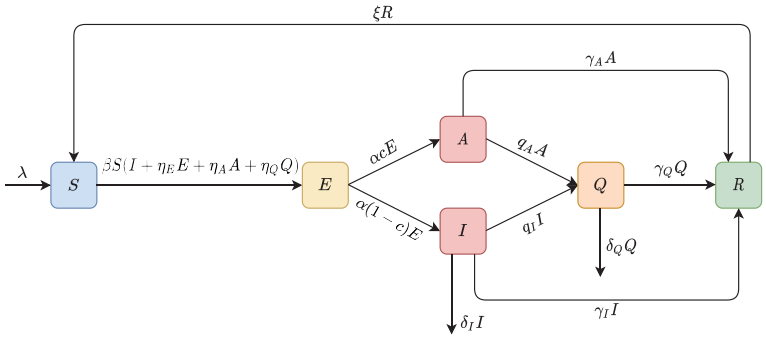


Figure 1. The compartment diagram of the full-model. λ is the constant recruitment rate; β is the transmission rate of I compartment; $\eta_i, i = E, A, Q$ are the multiple of the transmission rate of E , A and Q compartment relative to I compartment respectively; c is the proportion of asymptomatic infection; $\gamma_i, i = A, I, Q$ represent the recovery rate of A , I and Q compartment respectively; $\delta_i, i = I, Q$ represent the disease-induced death rate of I and Q compartment respectively; $\frac{1}{\alpha}$ is the incubation period; $q_i, i = A, I$ represent the detection rate of A and I compartment respectively; ξ is the rate of the recovered patients return to the susceptible population due to the weakening or disappearance of antibodies; ω is the natural mortality rate which is not shown in the figure.

The general ODE compartmental model corresponding to the compartment diagram is as follows.

$$\begin{cases} \dfrac{dS}{dt} = \lambda - \beta S(I + \eta_E E + \eta_Q Q + \eta_A A) + \xi R - \omega S, \\ \dfrac{dE}{dt} = \beta S(I + \eta_E E + \eta_Q Q + \eta_A A) - \alpha E - \omega E, \\ \dfrac{dA}{dt} = \alpha c E - q_A A - \gamma_A A - \omega A, \\ \dfrac{dI}{dt} = \alpha(1-c)E - q_I I - \gamma_I I - \delta_I I - \omega I, \\ \dfrac{dQ}{dt} = q_A A + q_I I - \gamma_Q Q - \delta_Q Q - \omega Q, \\ \dfrac{dR}{dt} = \gamma_A A + \gamma_I I + \gamma_Q Q - \xi R - \omega R. \end{cases} \tag{2.1}$$

2.1. Model selection

The full compartment model (2.1) can be simplified to various extents, based on different biological hypotheses, leading to different simplified models as shown in Table 1. We worry that "simplicity" may make us ignore some important factors describing the disease transmission. Therefore, we perform model selection to determine the best model that explains the data.

Table 1. List of models

Models	Assumptions	Biological meaning
Model 1	Full model	Distinguish between asymptomatic and symptomatic, suppose the recovery can be re-infected
Model 2	$\xi = 0$	Distinguish between asymptomatic and symptomatic, suppose the recovery will not be re-infected
Model 3	$c = 0, q_A = 0, \gamma_A = 0$	Not distinguish between asymptomatic and symptomatic, suppose the recovery can be re-infected
Model 4	$\xi = 0, c = 0, q_A = 0, \gamma_A = 0$	Not distinguish between asymptomatic and symptomatic, suppose the recovery will not be re-infected

Annotation. $I(t)$ of Model 1 and Model 2 denote the number of symptomatic individuals who are not diagnosed at time t. While $I(t)$ of Model 3 and Model 4 denote the number of infections who are not diagnosed at time t.

The Akaike Information Criterion (AIC) is a standard model selection criterion. But the criterion is only suitable when the number of time points is large in comparison to the number of parameters to be estimated. Therefore, the following modified AIC (AICc) should be used (see, [26], which is the AIC with a bias correction term

for a small sample size).

$$AICc = -2ln(L(\hat{\theta}_{MLE})) + 2K + \frac{2K(K+1)}{W-K-1},$$

where K is the number of unknown parameters, W is the sample size, L is the likelihood function. The smaller the AICc value of the model, the better that the model describes the data. In addition, we also calculate the BIC [25] to verify the reliability of model selection.

To obtain some important parameters of COVID-19 in the natural state (without any control measures), we select the number of remaining confirmed cases from February 29 to March 14, 2020 in the U.S. to fit the unknown parameters (i.e., Data1). These data reflect the acceleration of virus transmission dynamics in the population and avoid the influence of the prevention and control measures on fitting the key transmission parameters. Since some model parameters have been reasonably estimated in other literature, we fix these parameters as constant values (see Table 2). The key transmission parameters with local transmission characteristics (e.g., the transmission rate) are fitted.

Table 2. Definition and values of the constant parameter in the general ODE compartmental model

Parameter	Meaning	Value	Source
λ	The constant recruitment into the susceptible population	10424	Calculated
$1/\alpha$	The incubation period	5.2	[17]
γ_A	The recovery rate of A compartment	0.1397	[10]
γ_I	The recovery rate of I compartment	0.0698	[10]
δ_I	The disease-induced death rate of I	0.0412	[10]
γ_Q	The recovery rate of Q compartment	1/10.4	[9]
δ_Q	The disease-induced death rate of Q	0.015	[9]
ω	The natural mortality rate	3.425e-5	[32]
Q^0	The initial values of the confirmed cases reported	64	[7]
R^0	The initial values of the recovery	0	[7]

Annotation. $\lambda = N * \Lambda/365$, N denotes total population and Λ denotes birth rate.

In general, the observed data are always noisy. Method of maximum likelihood under Bayesian inference for dynamical systems can estimate the the data noise [23]. We assume that the probability model for the observed data is a normal distribution with mean given by $Q(t)$ and variance given by $\frac{1}{\tau}$. Since the variance $\frac{1}{\tau}$ of the noise distribution is unknown, we also take it as one of the parameters to be estimated. This paper uses the GWMCMC algorithm to calculate the posterior distributions for all estimated parameters and the initial values. We can obtain the best-fit values of the parameters and their 95% confidence intervals (95% CI) from the posterior distribution. All parameter estimates for four models are shown in Table 3. The results of model selection are shown in Table 4. As seen in Table 4, Model 4 has the smallest AICc, with AICc=174.5, which means that Model 4 is the most suitable model to explain the data. The BIC value also verifies this result. Therefore, considering the two factors of classifying undetected infections as asymptomatic or symptomatic and the recovered individuals can be re-infected or not, which can describe the transmission process of COVID-19 in more detail. However, it is not essential. Fig 2(a) and Fig 2(b) show the fitting curve and its corresponding 95% confidence interval of Model 4 on Data 1 (the number of confirmed cases).

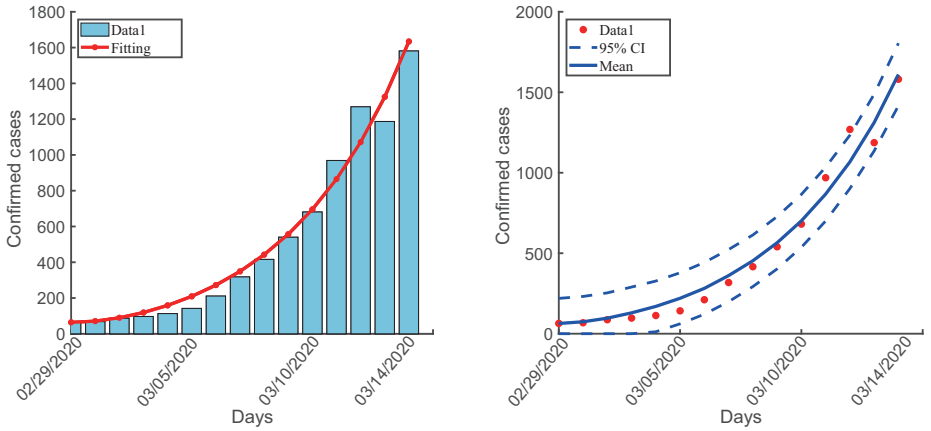
Table 3. Model parameters and their fitted values

Par	Model 1		Model 2		Model 3		Model 4	
	Best-fit value	95% CI	Best-fit value	95% CI	Best-fit value	95% CI	Best-fit value	95% CI
β	1.03e-09	(8.56, 11.4)e-10	1.27e-09	(1.21, 2.51)e-09	1.02e-09	(9.40, 15.9)e-10	1.11e-09	(1.06, 1.57)e-09
η_E	0.53	(0.50, 0.67)	0.46	(0.10, 0.54)	0.65	(0.25, 0.78)	0.59	(0.31, 0.60)
η_A	0.82	(0.59, 0.89)	0.64	(0.27, 0.68)	n/a	n/a	n/a	n/a
η_Q	0.02	(0.01, 0.17)	0.17	(0.03, 0.37)	0.27	(0.001, 0.44)	0.12	(0.01, 0.28)
c	0.40	(0.40, 0.56)	0.55	(0.42, 0.64)	n/a	n/a	n/a	n/a
q_A	0.002	(0.001, 0.004)	0.002	(0.001, 0.005)	n/a	n/a	n/a	n/a
q_I	0.13	(0.11, 0.18)	0.24	(0.10, 0.28)	0.09	(0.07, 0.11)	0.09	(0.08, 0.13)
ξ	0.49	(0.35, 0.73)	n/a	n/a	0.33	(0.01, 0.97)	n/a	n/a
τ	2.08e-04	(9.89, 28.8)e-05	2.48e-04	(9.54, 32.9)e-05	1.80e-04	(8.28, 25.8)e-05	1.81e-04	(8.57, 26.5)e-05
E^0	55	(53, 310)	58	(54, 505)	948	(765, 998)	927	(709, 999)
A^0	2084	(1581, 2199)	1773	(1199, 1998)	n/a	n/a	n/a	n/a
I^0	77	(50, 115)	51	(50, 107)	104	(100, 164)	76	(70, 142)

Annotation. $\frac{1}{\tau}$ is the variance of data noise.

Table 4. List of AICc and BIC for each model

Models	No. of parameter fitted	AICc	BIC	Maximum likelihood
Model 1	12	321.13	173.63	2.26e-31
Model 2	11	251.70	171.49	1.70e-31
Model 3	8	184.87	166.54	3.48e-32
Model 4	7	174.57	163.53	4.05e-32



(a) 2020.2.29-2020.3.14 (b) 95% credible interval

Figure 2. Fitting results. (a) Fitting of Model 4 to the data of COVID-19 in the United States from February 29 to March 14, 2020. (b) It's 95% confidence interval.

2.2. Stability analysis

In this section, we make a long-range forecast of Model 4 selected by AICc criterion and mainly explore its global convergence to the endemic state.

First, we get the basic reproduction number (denoted as $\mathcal{R}_0$) [6] of Model 4 as:

$$\begin{aligned}\mathcal{R}_0 =&\frac{\beta\eta_E\lambda}{\omega(\alpha+\omega)}+\frac{\beta\lambda\alpha}{\omega(q_I+\gamma_I+\delta_I+\omega)(\alpha+\omega)}\\&+\frac{\beta\eta_Q\lambda\alpha q_I}{\omega(q_I+\gamma_I+\delta_I+\omega)(\gamma_Q+\delta_Q+\omega)(\alpha+\omega)}.\end{aligned}\tag{2.2}$$

We have the following conclusions about the equilibrium state and stability of Model 4.

Theorem 2.1. *When $\mathcal{R}_0 \le 1$, the disease-free equilibrium $P_0 = (\frac{\lambda}{\omega}, 0, 0, 0, 0)$ of Model 4 is globally asymptotically stable.*

Proof. Consider the Lyapunov function

$$\mathcal{V} = \frac{\beta(\eta_E AB + \alpha B + \eta_Q\alpha q_I)}{ABC}E + \frac{\beta B + \beta\eta_Q q_I}{AB}I + \frac{\beta\eta_Q}{B}Q,\tag{2.3}$$

where $A = q_I + \gamma_I + \delta_I + \omega, B = \gamma_Q + \delta_Q + \omega, C = \alpha + \omega$, then

$$\begin{aligned}\mathcal{V}' =&\frac{\beta(\eta_E AB + \alpha B + \eta_Q\alpha q_I)}{ABC}[\beta S(I + \eta_E E + \eta_Q Q) - CE] + \frac{\beta B + \beta\eta_Q q_I}{AB}(\alpha E\\&- AI) + \frac{\beta\eta_Q}{B}(q_I I - BQ)\\=&\beta(I + \eta_E E + \eta_Q Q)\left[\frac{\beta(\eta_E AB + \alpha B + \eta_Q\alpha q_I)}{ABC}S - 1\right]\\\le&\beta(I + \eta_E E + \eta_Q Q)(\mathcal{R}_0 - 1).\end{aligned}$$

When $\mathcal{R}_0 \le 1$, the Lyapunov function $\mathcal{V}' \le 0$. Furthermore, $\mathcal{V}' = 0$ only if $E = I = Q = 0$ or $\mathcal{R}_0 = 1$. The maximum invariant set in $\{(S, E, I, Q, R) : \mathcal{V}' = 0\}$ is the singleton P_0. According to LaSalle-Lyapunov theory [15], the disease-free equilibrium P_0 is globally asymptotically stable when $\mathcal{R}_0 \le 1$. □

2.2.1. Long-range forecast: global convergence to the endemic state

In addition, when $\mathcal{R}_0 > 1$, Model 4 has an endemic equilibrium $P^* = (S^*, E^*, I^*, Q^*, R^*)$, where

$$\begin{aligned}&S^* = \frac{\lambda}{\omega}\frac{1}{\mathcal{R}_0},\quad E^* = \frac{\lambda}{C}\left(1 - \frac{1}{\mathcal{R}_0}\right),\quad I^* = \frac{\alpha\lambda}{AC}\left(1 - \frac{1}{\mathcal{R}_0}\right),\\&Q^* = \frac{q_I\alpha\lambda}{ABC}\left(1 - \frac{1}{\mathcal{R}_0}\right),\quad R^* = \frac{1}{\omega}(\frac{\alpha\lambda\gamma_I}{AC} + \frac{q_I\alpha\lambda\gamma_Q}{ABC})(1 - \frac{1}{R_0}).\end{aligned}$$

We have the following theorem for the stability of P^*:

Theorem 2.2. *When $\mathcal{R}_0 > 1$, the endemic state P^* is globally asymptotically stable for Model 4.*

Proof. Since the last variable of the system does not appear in the first four equations, we can only consider the first four variables and transform Model 4 into the following form:

$$\begin{cases} \dfrac{dS}{d\tau} = \tilde{\lambda} - \tilde{\beta}S(I + \eta_E E + \eta_Q Q) - S, \\ \dfrac{dE}{d\tau} = \tilde{\beta}S(I + \eta_E E + \eta_Q Q) - \tilde{\alpha}E - E, \\ \dfrac{dI}{d\tau} = \tilde{\alpha}E - \tilde{A}I - I, \\ \dfrac{dQ}{d\tau} = \tilde{q_I}I - \tilde{B}Q - Q, \end{cases} \tag{2.4}$$

where $\tilde{\lambda} = \dfrac{\lambda}{\omega}, \tilde{\beta} = \dfrac{\beta}{\omega}, \tilde{\alpha} = \dfrac{\alpha}{\omega}, \tilde{q_I} = \dfrac{q_I}{\omega}, \tilde{A} = \dfrac{q_I + \gamma_I + \delta_I}{\omega}, \tilde{B} = \dfrac{\gamma_Q + \delta_Q}{\omega}$. Let the right side of each of the four differential equations equal to zero in the system (2.4), obtaining the equations:

$$\begin{cases} \dfrac{\tilde{\lambda}}{S^*} - \tilde{\beta}(I^* + \eta_E E^* + \eta_Q Q^*) = 1, \\ \tilde{\beta}\dfrac{S^* I^*}{E^*} + \tilde{\beta}\eta_E S^* + \tilde{\beta}\eta_Q \dfrac{S^* Q^*}{E^*} = \tilde{\alpha} + 1, \\ \tilde{\alpha}\dfrac{E^*}{I^*} = \tilde{A} + 1, \\ \tilde{q_I}\dfrac{I^*}{Q^*} = \tilde{B} + 1. \end{cases} \tag{2.5}$$

Substituting (2.5) into (2.4), we have:

$$\begin{cases} \dfrac{dS}{d\tau} = S\left[\dfrac{\tilde{\lambda}}{S} - \dfrac{\tilde{\lambda}}{S^*} - \tilde{\beta}\eta_E(E - E^*) - \tilde{\beta}(I - I^*) - \tilde{\beta}\eta_Q(Q - Q^*)\right], \\ \dfrac{dE}{d\tau} = E\left[\tilde{\beta}\eta_E(S - S^*) + \tilde{\beta}\left(\dfrac{SI}{E} - \dfrac{S^* I^*}{E^*}\right) + \tilde{\beta}\eta_Q\left(\dfrac{SQ}{E} - \dfrac{S^* Q^*}{E^*}\right)\right], \\ \dfrac{dI}{d\tau} = I\left[\tilde{\alpha}\left(\dfrac{E}{I} - \dfrac{E^*}{I^*}\right)\right], \\ \dfrac{dQ}{d\tau} = Q\left[\tilde{q_I}\left(\dfrac{I}{Q} - \dfrac{I^*}{Q^*}\right)\right]. \end{cases} \tag{2.6}$$

Let $x = \dfrac{S}{S^*}, y = \dfrac{E}{E^*}, z = \dfrac{I}{I^*}, u = \dfrac{Q}{Q^*}$, then system (2.6) is equivalent to

$$\begin{cases} x' = x\left[\dfrac{\tilde{\lambda}}{S^*}\left(\dfrac{1}{x} - 1\right) - \tilde{\beta}\eta_E E^*(y - 1) - \tilde{\beta}I^*(z - 1) - \tilde{\beta}\eta_Q Q^*(u - 1)\right], \\ y' = y\left[\tilde{\beta}\eta_E S^*(x - 1) + \tilde{\beta}\dfrac{S^* I^*}{E^*}\left(\dfrac{xz}{y} - 1\right) + \tilde{\beta}\eta_Q \dfrac{S^* Q^*}{E^*}\left(\dfrac{xu}{y} - 1\right)\right], \\ z' = z\left[\tilde{\alpha}\dfrac{E^*}{I^*}\left(\dfrac{y}{z} - 1\right)\right], \\ u' = u\left[\tilde{q_I}\dfrac{I^*}{Q^*}\left(\dfrac{z}{u} - 1\right)\right]. \end{cases} \tag{2.7}$$

Consider the Lyapunov function

$$V = S^*\left(x - 1 - \ln x\right) + E^*\left(y - 1 - \ln y\right) + \frac{\tilde{\beta}\eta_Q S^* I^* Q^* + \tilde{\beta} S^* (I^*)^2}{\tilde{\alpha} E^*}\left(z - 1 - \ln z\right) + \frac{\tilde{\beta}\eta_Q S^* (Q^*)^2}{\tilde{q_I} I^*}\left(u - 1 - \ln u\right), \tag{2.8}$$

then

$$\begin{aligned} V' =& S^*\left(1 - \frac{1}{x}\right)x' + E^*\left(1 - \frac{1}{y}\right)y' + \frac{\tilde{\beta}\eta_Q S^* I^* Q^* + \tilde{\beta} S^* (I^*)^2}{\tilde{\alpha} E^*}\left(1 - \frac{1}{z}\right)z' \\ &+ \frac{\tilde{\beta}\eta_Q S^* (Q^*)^2}{\tilde{q_I} I^*}\left(1 - \frac{1}{u}\right)u' \\ =& S^*(x-1)\left[\frac{\tilde{\lambda}}{S^*}\left(\frac{1}{x} - 1\right) - \tilde{\beta}\eta_E E^*(y-1) - \tilde{\beta} I^*(z-1) - \tilde{\beta}\eta_Q Q^*(u-1)\right] \\ &+ E^*(y-1)\left[\tilde{\beta}\eta_E S^*(x-1) + \tilde{\beta}\frac{S^* I^*}{E^*}\left(\frac{xz}{y} - 1\right) + \tilde{\beta}\eta_Q \frac{S^* Q^*}{E^*}\left(\frac{xu}{y} - 1\right)\right] \\ &+ \frac{\tilde{\beta}\eta_Q S^* I^* Q^* + \tilde{\beta} S^* (I^*)^2}{\tilde{\alpha} E^*}(z-1)\left[\tilde{\alpha}\frac{E^*}{I^*}\left(\frac{y}{z} - 1\right)\right] \\ &+ \frac{\tilde{\beta}\eta_Q S^* (Q^*)^2}{\tilde{q_I} I^*}(u-1)\left[\tilde{q_I}\frac{I^*}{Q^*}\left(\frac{z}{u} - 1\right)\right] \\ =& \tilde{\beta} S^* I^*\left(3 - \frac{xz}{y} - \frac{y}{z} - \frac{1}{x}\right) + \tilde{\beta}\eta_Q S^* Q^*\left(4 - \frac{xu}{y} - \frac{y}{z} - \frac{z}{u} - \frac{1}{x}\right) \\ &+ \left[\tilde{\lambda} - (\tilde{\beta} S^* I^* + \tilde{\beta}\eta_Q S^* Q^*)\right]\left(2 - x - \frac{1}{x}\right), \end{aligned}$$

where $\tilde{\lambda} - \tilde{\beta} S^* I^* - \tilde{\beta}\eta_Q S^* Q^* = S^* + \tilde{\beta} S^* E^* > 0$, and

$$\begin{aligned} &\frac{xz}{y} + \frac{y}{z} + \frac{1}{x} \le 3, \\ &\frac{xu}{y} + \frac{y}{z} + \frac{z}{u} + \frac{1}{x} \le 4, \\ &x + \frac{1}{x} \le 2. \end{aligned}$$

Therefore, $V' \le 0$. $V' = 0$ only if $x = 1, y = z = u$, where S, E, I, Q satisfy the set:

$$M = \left\{(S, E, I, Q) | S = S^*, \frac{E}{E^*} = \frac{I}{I^*} = \frac{Q}{Q^*}\right\}.$$

Since the equilibrium point (S^*, E^*, I^*, Q^*) is the unique invariant set of the system (2.4), the endemic equilibrium P^* is globally asymptotically stable when $\mathcal{R}_0 > 1$. □

The global convergence to the endemic state means that COVID-19 will become endemic in the United States in the absence of strict control measures. Simulations below will also show that the COVID-19 infection in the U.S. will undergo damped transit oscillations before the final endemic state is reached. We should anticipate future and multiple outbreaks after the vaccination rollout.

3. Identifiability analysis

This section aims to study the identifiability of fitted parameters in Model 4. We will study its structural identifiability and practical identifiability separately.

3.1. Structural identifiability analysis

This section analyzes whether each estimated parameter value in Model 4 can be uniquely determined under the ideal condition with perfect, noise-free data. The model is not structurally identifiable if two or more parameter sets can lead to the same observational output. Under this situation, the estimation of the parameters in the model might not be unique and thus the prediction from the model will be unreliable. About structural identifiability, Miao et al. [22] gave a clear definition like the following:

Definition 3.1. A general dynamic system can be expressed as follows:

$$\begin{aligned}\boldsymbol{x'}(t) &= \boldsymbol{f}(\boldsymbol{x}(t), \boldsymbol{p}),\\ \boldsymbol{y}(t) &= g(\boldsymbol{x}(t), \boldsymbol{p}),\end{aligned}$$

where $\boldsymbol{x}(t) \in R^m$ is a vector of state variables, $\boldsymbol{y}(t) \in R^d$ is the measurement or output vector, the parameter vector $\boldsymbol{p} \in R^q$. A parameter set $\boldsymbol{p}$ is called structurally (or uniquely) identifiable if for every $\boldsymbol{q}$ in the parameter space, the equation $g(\boldsymbol{x}(t), \boldsymbol{p}) = g(\boldsymbol{x}(t), \boldsymbol{q})$ holds if and only if $\boldsymbol{p} = \boldsymbol{q}$.
Any unequal parameter set yields different observations and hence the corresponding noise-free data are distinct.

We will use the differential algebra approach to study the structural identifiability of Model 4. This method builds upon the derivation of the input-output equation, which contains all the structural identifiability information of the model. For the structural identifiability of Model 4, we obtain the following theorem:

Theorem 3.1. *When parameters* λ, $\frac{1}{\alpha}$, $\delta_i, i = I, Q$, $\gamma_i, i = A, I, Q$ *are fixed, the unknown parameter set* $\boldsymbol{p} = [\beta \quad \eta_E \quad \eta_Q \quad q_I]$ *of Model 4 is structurally identifiable from Data 1 (the number of confirmed cases observed).*

Proof. The equation of variable R in Model 4 is unnecessary for determining the model behavior and so is omitted. From the equation of $Q(t)$ (the fifth equation), we can get

$$I = \frac{Q' + (\gamma_Q + \delta_Q + \omega)Q}{q_I}. \tag{3.1}$$

Taking the derivative of (3.1), we can get

$$I' = \frac{Q'' + (\gamma_Q + \delta_Q + \omega)Q'}{q_I}. \tag{3.2}$$

Plugging (3.1) and (3.2) into the equation about the variable I in Model 4 and then take the derivative of the equation yielding:

$$E = \frac{Q'' + (\gamma_Q + \delta_Q + q_I + \gamma_I + \delta_I + 2\omega)Q' + (q_I + \gamma_I + \delta_I + \omega)(\gamma_Q + \delta_Q + \omega)Q}{\alpha q_I}. \tag{3.3}$$

$$E' = \frac{Q^{(3)} + (\gamma_Q + \delta_Q + q_I + \gamma_I + \delta_I + 2\omega)Q'' + (q_I + \gamma_I + \delta_I + \omega)(\gamma_Q + \delta_Q + \omega)Q'}{\alpha q_I}. \tag{3.4}$$

Plugging (3.1), (3.3) and (3.4) into the model E equation of Model 4 and take the derivative yielding:

$$S = \frac{Q^{(3)} + DQ'' + C_1Q' + FQ}{\beta(\eta_E Q'' + A_1Q' + B_1Q)}. \tag{3.5}$$

$$\begin{aligned} S' =& \frac{Q^{(4)} + DQ^{(3)} + C_1Q'' + FQ'}{\beta(\eta_E Q'' + A_1Q' + B_1Q)} \\ &+ \frac{\beta(Q^{(3)} + DQ'' + C_1Q' + FQ)(\eta_E Q^{(3)} + A_1Q'' + B_1Q')}{[\beta(\eta_E Q'' + A_1Q' + B_1Q)]^2}. \end{aligned} \tag{3.6}$$

Where

$$\begin{aligned} &A_1 = \alpha + (q_I + \gamma_I + \delta_I + \omega)\eta_E + (\gamma_Q + \delta_Q + \omega)\eta_E. \\ &B_1 = \alpha(\gamma_Q + \delta_Q + \omega) + \alpha\eta_Q q_I + (q_I + \gamma_I + \delta_I + \omega)(\gamma_Q + \delta_Q + \omega)\eta_E. \\ &C_1 = (\alpha+\omega)(\gamma_Q+\delta_Q+\omega)+(q_I+\gamma_I+\delta_I+\omega)(\gamma_Q+\delta_Q+\omega)+(\alpha+\omega)(q_I+\gamma_I+\delta_I+\omega). \\ &D = \alpha + q_I + \gamma_I + \delta_I + \gamma_Q + \delta_Q + 3\omega. \\ &F = (\alpha + \omega)(q_I + \gamma_I + \delta_I + \omega)(\gamma_Q + \delta_Q + \omega). \end{aligned}$$

Plugging (3.1), (3.3), (3.5) and (3.6) into the $S(t)$ equation of Model 4. Then we obtain the input-output equation as follows:

$$\begin{aligned} &\alpha\eta_E q_I Q^{(4)}Q'' + \alpha q_I A_1 Q^{(4)}Q' + \alpha q_I B_1 Q^{(4)}Q - \alpha\eta_E q_I (Q^{(3)})^2 + \beta\eta_E^2 Q^{(3)}(Q'')^2 \\ &+ 2\beta\eta_E A_1 Q^{(3)}Q''Q' + 2\beta\eta_E B_1 Q^{(3)}Q''Q + \alpha q_I(\eta_E\omega - A_1)Q^{(3)}Q'' + \beta A_1^2 Q^{(3)}(Q')^2 \\ &+ 2\beta A_1 B_1 Q^{(3)}Q'Q + \alpha q_I[A_1(D+\omega) - \eta_E C_1 - B_1]Q^{(3)}Q' + \beta B_1^2 Q^{(3)}Q^2 \\ &+ \alpha q_I[B_1(D+\omega) - \eta_E F]Q^{(3)}Q + \beta\eta_E^2 D(Q'')^3 + \beta\eta_E(2A_1D + \eta_E C_1)(Q'')^2Q' \\ &+ \beta\eta_E(2B_1D + \eta_E F)(Q'')^2Q + \alpha q_I[D(\eta_E\omega - A_1) + \eta_E C_1 - \lambda\beta\eta_E^2](Q'')^2 \\ &+ \beta(A_1D + 2\eta_E C_1)A_1Q''(Q')^2 + 2\beta[A_1B_1(D+\eta_E) + \eta_E B_1 C_1 + \eta_E A_1 F]Q''Q'Q \\ &+ \alpha q_I[-B_1D + (\omega D - 2\lambda\beta\eta_E)A_1 + \eta_E\omega C_1 + \eta_E F]Q''Q' + \beta(B_1D \\ &+ 2\eta_E F)B_1Q''Q^2 + \alpha q_I[(\omega D + C_1 - 2\lambda\beta\eta_E)B_1 + (\eta_E\omega - A_1)F]Q''Q \\ &+ \beta A_1^2 C(Q')^3 + \beta(2A_1B_1C_1 + A_1^2F)(Q')^2Q + \alpha q_I[-B_1C_1 + (\omega C_1 + F \\ &- \lambda\beta A_1)A_1](Q')^2 + \beta(B_1C_1 + 2A_1F)B_1Q'Q^2 + \alpha q_I[\omega(C_1+F) - 2\lambda\beta A_1]B_1Q'Q \\ &+ \beta B_1^2 FQ^3 + \alpha q_I(\omega F - \lambda\beta B_1)B_1Q^2 = 0. \end{aligned} \tag{3.7}$$

We need to get the normalized input-output equation from (3.7). Although the normalized input-output equation coefficients are very complex, there are only four unknown parameters. Suppose that another parameter set $\boldsymbol{q} = [q_1 \quad q_2 \quad q_3 \quad q_4]$ can produce the same output, we can choose four coefficients of $Q^{(4)}Q'$, $Q^{(4)}Q$, $Q^{(3)}(Q'')^2$ and $(Q'')^3$ to form the following equation group by using the injectivity

of the coefficients of input-output equations [8]:

$$\begin{cases} \dfrac{\alpha q_I(\alpha+A\eta_E+B\eta_E)}{\alpha\eta_E q_I}=\dfrac{\alpha q_4[\alpha+(q_4+\gamma_I+\delta_I+\omega)q_2+Bq_2]}{\alpha q_2 q_4}, \quad \dfrac{\beta\eta_E^2}{\alpha\eta_E q_I}=\dfrac{q_1 q_2^2}{\alpha q_2 q_4}, \\ \dfrac{\beta\eta_E^2\alpha+A+B+\omega}{\alpha\eta_E q_I}=\dfrac{q_1 q_2^2\alpha+q_4+\gamma_I+\delta_I+\omega+B+\omega}{\alpha q_2 q_4}, \\ \dfrac{\alpha q_I[\alpha B+\alpha\eta_Q q_I+\eta_E AB]}{\alpha\eta_E q_I}=\dfrac{\alpha q_4[\alpha B+\alpha q_3 q_4+(q_4+\gamma_I+\delta_I+\omega)q_2 B]}{\alpha q_2 q_4}, \end{cases}$$

where $A=q_I+\gamma_I+\delta_I+\omega$,$B=\gamma_Q+\delta_Q+\omega$. Solving this equation group, we get

$$\beta=q_1,\eta_E=q_2,\eta_Q=q_3,q_I=q_4.$$

Therefore, Model 4 is structurally identifiable when λ, $\frac{1}{\alpha}$, $\delta_i,(i=I,Q)$, $\gamma_i,(i=A,I,Q)$ are fixed. □

More data sets (if any) can be considered to fit all parameters in the model, rather than only four parameters with local epidemic characteristics. However, we also find that the model is not structurally identifiable for all the parameters in the model. Even if multiple data sets are used, the fitting values of the parameters are unreliable. In practice, we fix some parameters relevant to the disease characteristics that have been estimated reliably from statistics (perhaps from data from countries that experienced the pandemic earlier, such as China and Italy). We further use available data in the study region (e.g., the U.S.) to identify the strength of public health interventions and local transmission characteristics.

3.2. Practical identifiability analysis

The parameters fitted in Model 4 are structurally identifiable. It is also a prerequisite to obtaining reliable parameter estimation from observational data. The structural identifiability depends on the assumption that the data are noise-free. However, the statistics data we got are embedded with noise. Therefore, a key problem when fitting a model to data is the influence of noise on parameter estimation and model identifiability.

First, we show the robustness of parameter estimation to explore the influence of different noise distributions. We test several common error models in our simulations: Poisson, Gaussian error (with standard deviation equal to 10% of the mean), and negative binomial (taking into account over-dispersion, letting the variance equal to five times the mean and the variance equal to fifty times the mean). For each error model, the parameter values fitted by Model 4 in the previous section are taken as the "real value" and brought into the model for prediction. We simulate 100 realizations by adding noise to the number of confirmed cases in the prediction results. Then, we use the GWMCMC algorithm to re-estimate the model parameters for each realization of the data. The resulting parameter estimates under the four different noise types are summarized in Table 5. The results show that the estimated results are similar for all distributions and estimation methods: most of the parameter estimations are close to the true parameter values. This indicates that the parameters estimated by using the GWMCMC algorithm are robust to the distribution of the errors of data.

Next, to quantify the practical identifiability of the parameters estimated, we take the Gaussian error distribution model as an example to carry on the further

Table 5. Parameter estimates made from 100 simulated data sets assuming different distributions (Poisson distribution, Gaussian distribution with $\sigma_0 = 10\%$, Negative binomial distribution with variance equal to 5 times the mean and variance equal to 50 times the mean) using maximum likelihood estimation, and true parameters are as in Table. 3 (Model 4)

Parameter	True value	Poisson	Gaussian (σ_0=10%)	Negative biomial (5 Times)	Negative biomial (50 Times)
β	1.11e-09	1.0968e-09	1.09876e-09	1.1131e-09	1.1219e-09
η_E	0.59	0.6023	0.60134	0.5877	0.5767
η_Q	0.12	0.1200	0.11982	0.1188	0.1197
q_I	0.09	0.0905	0.09187	0.0902	0.0894
$\mathcal{R}_0$	2.89	2.8615	2.8685	2.8786	2.8925

Annotation. True values are taken from the best-fit value of Model 4 in Table 3.

inquisition. To realize this, we generate 1000 simulated data sets under different measurement error levels (0%, 5%, 10%, 20%). We still performed the GWMCMC algorithm to estimate parameters.

The Monte Carlo simulation steps we used are as in [22]. Here, we take the best-fit value of Model 4 in Table 3 as the true parameter set $\boldsymbol{p}_0$.

(1) Solve the epidemiological model numerically with the true parameters $\boldsymbol{p}_0$ and obtain the output vector $Q(t_k)$ at the discrete data time points $\{t_k\}_{k=1}^n$.

(2) Generate $M = 1000$ residual vectors $\hat{\boldsymbol{r}}_i$ drawn from a normal distribution whose mean is the output vector computed in step (1) and standard deviation is the $\sigma_0\%$ of the mean.

(3) Fit the Model 4 to each of the M simulated data sets by maximizing the likelihood function using the Goodman and Weare affine ensemble Markov chain Monte Carlo algorithm to estimate the parameter set $\boldsymbol{p}_j$ for $j = 1, 2, ..., M$.

(4) Calculate the average relative estimation error (ARE) for each parameter in the set $\boldsymbol{p}$ by

$$ARE(p^{(k)}) = \frac{1}{M}\sum_{j=1}^{M}\frac{|\hat{p}_0^{(k)} - p_j^{(k)}|}{\hat{p}_0^{(k)}} \times 100\%,$$

where $p^{(k)}$ is the kth parameter in the set $\boldsymbol{p}$, $\hat{p}_0^{(k)}$ is the kth parameter in the true parameter set $\boldsymbol{p}_0$, $p_j^{(k)}$ is the kth parameter in the estimated parameter set $\boldsymbol{p}_j$ from the jth simulation data set, and M is the total number of simulation runs.

(5) Repeat steps 1 through 4 with increasing level of noise, that is take $\sigma_0 = 0, 5, 10, 20\%$.

The ARE can be used to evaluate whether or not the estimated value of each parameter is acceptable. For minimal measurement error, the estimated value of the parameter should be close to the true values and the ARE should be chose to 0. When the measurement error increases, the AREs of the parameter estimation should also increase. In the practical identifiability analysis, if the estimated values of the parameter are not sensitive to measurement errors, we say that the estimated parameter is practically identifiable. Otherwise, if the ARE of the estimated parameter is large even for a small measurement error, we claim that it is practically unidentifiable. However, there is no clear-cut rule on the cut-off of the AREs before they are claimed to be "unacceptable" for a particular problem. Thus, practical identifiability depends on the underlying problem and judgment of

the investigators. In addition, when diverse statistical algorithms are used to estimate parameters, the values of ARE may be different [22]. Therefore, we take a simplistic approach: if the ARE of that parameter is greater than the measurement error level, the parameter is practically unidentifiable [29].

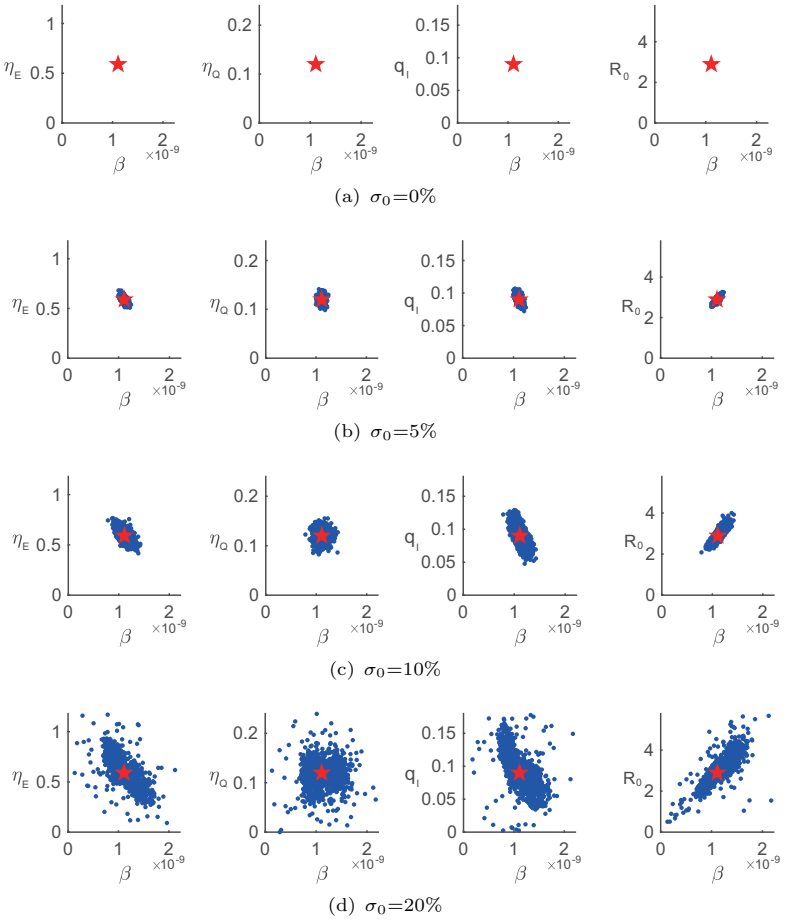
We generate 1000 simulated data sets under different measurement error levels (0%, 5%, 10%, 20%), and perform a GWMCMC algorithm to estimate the unknown parameters of Model 4 for each set of data. The AREs of all unknown parameters of Model 4 for four measurement error levels (0%, 5%, 10%, 20%) are reported in Table 6. From Table 6 we can see that for the situation of no measurement error ($\sigma_0 = 0\%$), all the four parameter can be well identified (the maximum ARE is 0.84%, very close to 0), which confirms our structural identifiability analysis in section 3.1. This also indicates that the reliability of the parameter estimation method. With the increase of measurement error levels, the AREs of four parameters increase gradually. In all cases, the AREs of η_Q are less than the corresponding measurement error levels, which means η_Q is practically identifiable. But when the measurement error increases to 20%, β and η_E change from identifiable to unidentifiable. This means that the two parameters are practically identifiable for medium measurement errors, but become unidentifiable for larger errors. On the other hand, Table 3 shows the variance of Data 1's error ($\frac{1}{\tau}$) estimated through the GWMCMC to be 5524, which is equivalent to the noise level σ_0 equal to about 10%. In this sense, it is reasonable to assume that parameters β and η_E are also practically identifiable in our research. Unfortunately, the parameter q_I is hard to be identified even for the 5% measurement error level.

Table 6. Practical identifiability analysis of parameters of Model 4 by the affine invariant ensemble Markov chain Monte Carlo simulations for measurement error levels $\sigma_0 = 0, 5, 10$ and 20%

Parameter	ARE 0%	ARE 5%	ARE 10%	ARE 20%
β	0.45%	2.69%	7.01%	20.90%
η_E	0.72%	3.50%	8.00%	23.29%
η_Q	0.84%	4.15%	8.25%	19.75%
q_I	0.77%	5.13%	13.08%	28.51%

To vividly describe each parameter's identifiability, we draw the scatter plots of parameter estimation at different noise levels (Fig 3), where red stars represent real parameter values. It can be seen from Fig 3 that as the error level increases, the scatter plots become more and more dispersed. When the error level is not more than 10 %, except the parameter q_I, the other three parameters (β, η_E and η_Q) are gathered near the real value. In addition, we find that for all error levels, the parameter β is positively correlated with the basic reproduction number $\mathcal{R}_0$.

In addition, we consider the statistical information of the basic reproduction number $\mathcal{R}_0$ in 1000 simulations. We find that when the error level $\sigma_0 = 10\%$, the upper quartile $Q_1 = 2.66$ and the lower quartile $Q_3 = 3.07$. When the error level increases to $\sigma_0 = 20\%$, the upper quartile of $\mathcal{R}_0$ is $Q_1 = 2.40$, and the lower quartile is $Q_3 = 3.46$. This shows that when the error level is small, $\mathcal{R}_0$ will be near the real value with greater probability, which is consistent with the frequency distribution histograms of $\mathcal{R}_0$ shown in Fig 4.



(a) σ_0=0%

(b) σ_0=5%

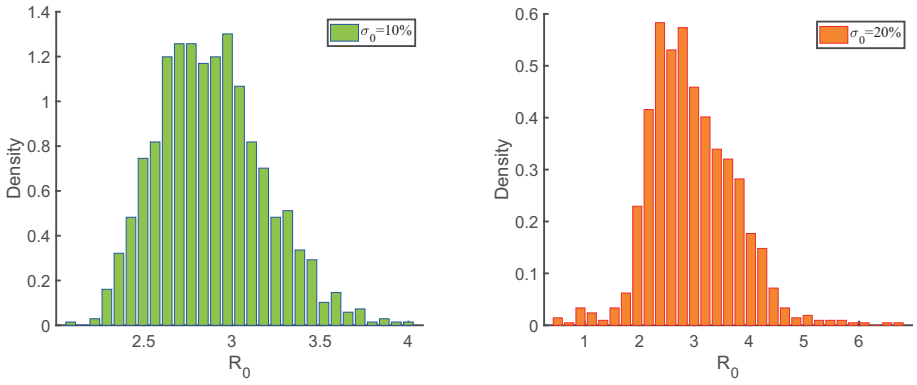
(c) σ_0=10%

(d) σ_0=20%

Figure 3. The parameter estimates of the Model 4 for 1000 synthetic data generated by Gaussian noise. True parameters are indicated by red stars.

4. Transit behaviors and long-range forecasts

Model 4 is selected by AICc to describe the early epidemic of the United States, which does not consider the impact of individual behavior changes and intervention measures on the last epidemic (including wearing masks, vaccination, etc). In this section, based on Model 4, we add the time-dependent vaccination effect and the exponential form of social distance term to simulate and predict the scale of the recent epidemic in the United States. In literature [31], the authors use this exponential function to describe the impact of changes in social distancing on the epidemic. A similar functional form is used in our model. Let κ be the social distance parameter. $\kappa > 0$ indicates that the social distance inhibits the spread of the epidemic. Otherwise, it promotes the spread of the epidemic. It is worth mentioning that from March 15, 2021 to May 10, 2021, the trend of the confirmed cases has a significant turning point on April 6, due to the different prevention and control measures taken in this period. Let t^* denote the turning point and let two Heaviside functions ($H_\epsilon(t) = \epsilon + H(t-t^*)\epsilon_1$ and $H_\kappa(t) = \kappa + H(t-t^*)\kappa_1$) embody



(a) σ_0=10% (b) σ_0=20%

Figure 4. The frequency distribution histograms of $\mathcal{R}_0$ from 1000 simulations for measurement error levels $\sigma_0 = 10\%$ and $\sigma_0 = 20\%$.

policy changes, where

$$H(t-t^*) = \begin{cases} 0, & t < t^* \\ 1, & t > t^* \end{cases}$$

So when $t < t^*$, the social distance parameter is κ and the vaccination effect parameter is ϵ. When $t > t^*$, the social distance parameter is $\kappa + \kappa_1$ and the vaccination effect parameter is $\epsilon + \epsilon_1$. Then the dynamic model during this period is as follows:

$$\begin{cases} \dfrac{dS}{dt} = \lambda - \beta(1 - [\epsilon + H(t-t^*)\epsilon_1])(I + \eta_E E + \eta_Q Q)S\Lambda(t) - \omega S, \\ \dfrac{dE}{dt} = \beta(1 - [\epsilon + H(t-t^*)\epsilon_1])(I + \eta_E E + \eta_Q Q)S\Lambda(t) - (\alpha + \omega)E, \\ \dfrac{dI}{dt} = \alpha E - q_I I - \gamma_I I - \delta_I I - \omega I, \\ \dfrac{dQ}{dt} = q_I I - \gamma_Q Q - \delta_Q Q - \omega Q, \\ \dfrac{dR}{dt} = \gamma_I I + \gamma_Q Q - \omega R, \end{cases} \tag{4.1}$$

where $\Lambda(t) = e^{-[\kappa + H(t-t^*)\kappa_1]\frac{E+I+Q}{S+E+I+Q+R}}$.

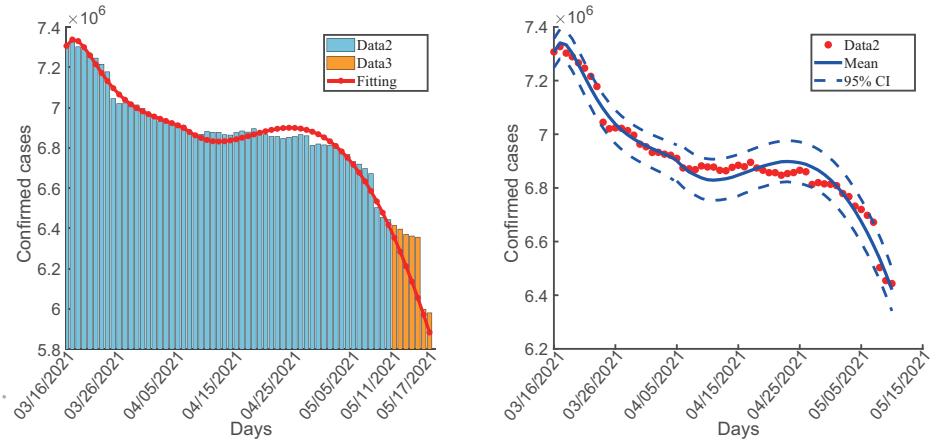
We use the number of confirmed cases from March 15 to May 10, 2021 in the U.S. (marked as Data 2) to estimate the social distance parameters κ and κ_1, the vaccination effect parameters ϵ and ϵ_1 in the system (4.1). The detection rate q_I is practically unidentifiable in Model 4, and more importantly, the detection rate may change in a different period, so we will also estimate q_I under Data 2. Since the other three parameters (β, η_E and η_Q) that reflect the characteristics of the virus itself are practically identifiable, we keep their values unchanged as in Table 3.

Estimates of these parameters and corresponding 95% confidence intervals are given in Table 7. Data 2 (the blue bar chart) and the best fitting curve (the red curve) are shown in Fig 5(a). The orange bar chart in Fig 5(a) represents the number of confirmed cases from May 11 to May 17, 2021 (denoted as Data 3). We

use it to verify the reliability of the results in Table 7. The 95 % confidence interval corresponding to the best fitting curve is shown in Fig 5(b). Here, $\kappa_1 = -17.21 < 0$ indicates that the social distance of the U.S. after the turning point t^* (April 6, 2021) is much smaller than that before t^*. On the other hand, $\epsilon_1 = 0.51 > 0$ indicates that the vaccination effect after t^* is much larger than that before t^*, which may improve the vaccine coverage.

Table 7. Definition and fitted values of parameters in the model (4.1).

Parameter	Meaning	Best-fit Value	95% confidence interval
ϵ	The effect of vaccination	0.12	(0.01, 0.26)
ϵ_1	The effect of vaccination	0.51	(0.49, 0.52)
κ	The social distance rate	11.24	(6.94, 12.35)
κ_1	The social distance rate	-17.21	(-17.55, -16.18)
q_I	The detection rate	0.17	(0.17, 0.21)
E^0	The initial value of E	5.64e+06	(5.35e+06, 5.65e+06)
I^0	The initial value of I	4.99e+06	(4.23e+06, 4.99e+06)
R^0	The initial value of R	2.61e+07	(2.28e+07, 4.45e+07)
τ_1	$\frac{1}{\tau_1}$ is the variance of data noise from March 15 to April 5, 2021	1.67e-09	(1.01e-09, 2.45e-09)
τ_2	$\frac{1}{\tau_2}$ is the variance of data noise from April 6 to May 10, 2021	7.61e-10	(5.01e-10, 9.83e-10)

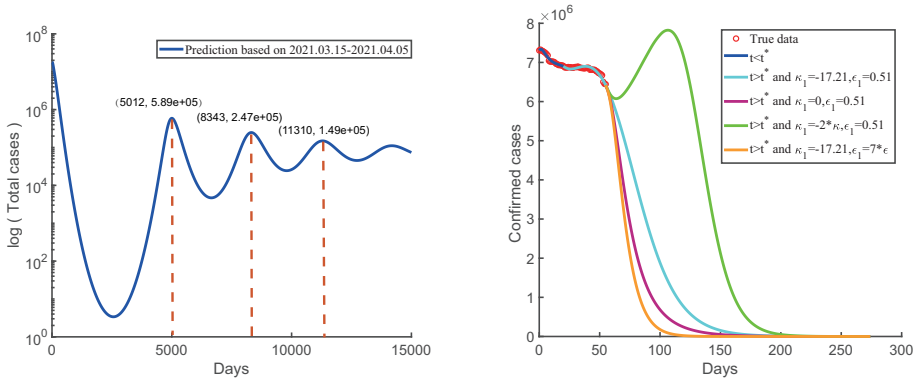


(a) 2021.3.16-2021.5.17 (b) 95% credible interval

Figure 5. Fitting results. (a)Piecewise fit of the model (4.1) to the data of COVID-19 in the United States from March 15 to May 10, 2021. (b)95% confidence interval

We find that if the United States maintains the same prevention and control strategy as that before the turning point t^* (i.e $\kappa = 11.24, \epsilon = 0.12$), the COVID-19 epidemic may break out again after several years (Fig 6(a)). Since the numerical span on the ordinate is big, we show the figure on the logarithmic scale against time so that the dynamic characteristic of the curve can be displayed more clearly. The longitudinal axis in Fig 6(a) represents the logarithm of the number of all cases in the United States. In addition, we also mark the peak value and corresponding time point of the possible outbreak of COVID-19 in the United States. The peaks

of each outbreak will gradually weaken over time. However, there will still be a large number of infections, and it will eventually become a local epidemic in the United States.



(a) (b)

Figure 6. Prediction in different situations.(a) The prediction of total cases with $\kappa = 11.24, \epsilon = 0.12$ and $t < t^*$. (b) The prediction of the number of confirmed cases with different values of κ_1 and ϵ_1.

We study the impact of several different social distances and vaccination on the epidemic trend (Fig 6(b)). All the curves in the figure are under the condition that $\kappa = 11.24, \epsilon = 0.12$. The longitudinal axis in Fig 6(b) represents the number of confirmed cases reported in the U.S. For example, the dark-blue line is the best fitting curve for March 15, 2021 to April 5, 2021. The sky-blue line is the best fitting curve for April 6, 2021 to May 10, 2021 and the forecast curve after that. Compared with the purple line and the sky-blue line we find that a reduction in social distancing would increase the number of confirmed cases, while maintaining the vaccine's effectiveness. As social distancing continues to decline, there will be an outbreak of the diagnoses (the green line). It suggests that proper social distancing is necessary until the outbreak is truly over. In addition, if maintain the social distancing between April 5, 2021 and May 10, 2021 and only increase the vaccine effectiveness, the number of diagnosed will drop rapidly (the yellow line). These data show that maintaining appropriate social distancing and strengthening vaccination are both effective measures to control the outbreak.

5. Discussion

If the parameters of the model can be properly estimated, the prediction conclusions will be more reliable. Structural identifiability issues for parameter estimation have been studied for some biological systems. In this paper, we establish a dynamic model that strictly distinguishes between confirmed and unconfirmed infections to describe the spread of COVID-19 in the U.S. We have discussed four models in this manuscript. The significant differences include: (1) whether the undetected infections are divided into asymptomatic or symptomatic. (2) whether the recovery will be infected again. All of the four models are fitted with the number of confirmed cases in the U.S. The result shows that the most suitable model is the one that ignores the difference between asymptomatic and symptomatic, and assumes that

recovered patients will not be re-infected. In other words, it is relatively optimal among the four models for interpreting the confirmed case data.

Choosing a model suitable for the data is only the first step. It is also very important to uniquely determine the model's unknown parameters from the actual data. We need to determine which parameters can be uniquely determined by these data, while other parameters cannot. Therefore, parameters' identifiability analyses must be performed before any statistical method is applied to estimate the unknown parameters from the experimental data. We prove that four unknown parameters of the selected model can be uniquely determined with noise-free data. If the model is structurally unidentifiable, there is no need for further study. However, although the model is structurally identifiable, these parameters may be not practically identifiable. At this moment, a suitable algorithm is critical to get optimal estimations of parameter values. Monte Carlo simulation is an important tool to validate the identifiability analysis results, perform sensitivity analyses for model parameters, and evaluate parameter estimation methods. In this manuscript, we use the GWMCMC algorithm to analyze the practical identifiability of the selected model. These results show that some parameters are identifiable at a negligible noise level. However, as the noise of data increases, the parameter identifiability weakens.

More data sets (if any) can be considered to fit all parameters in the model, rather than only four parameters with local epidemic characteristics. However, we find that the model is not structurally identifiable for all the parameters in the model. That is, even if multiple data sets are used, the best-fitted parameter values are unreliable. The structural unidentifiability of the model may be due to the structure of the model itself. In practice, we can fix some parameters that have been estimated reliably from statistics, and only fit those "difficult-to-obtain" parameters by traditional statistical methods.

The study suggests that if the officials in the U.S. do not respond to the outbreak with some degree of containment, the novel coronavirus will remain volatile in the United States and eventually become endemic. We also find that strengthening social distancing policy and improving vaccination effectiveness can accelerate the end of the epidemic. Otherwise, the epidemic will outbreak again.

Although the model can explain and predict the transmission characteristics of the COVID-19 in the U.S., it cannot study more complex outbreaks, such as the transmission of mutant strains, due to its relatively simple structure. We will focus on the relevant research in the future. In conclusion, we suggest that the selection of models and the identification analysis of parameters should be considered as one of the important research directions.

References

[1] H. Akaike, *Information theory and an extension of the maximum likelihood principle*, in *Selected papers of hirotugu akaike*, Springer, 1998, 199–213.

[2] G. Bellu, M. P. Saccomani, S. Audoly and L. D'Angiò, *Daisy: A new software tool to test global identifiability of biological and physiological systems*, Computer methods and programs in biomedicine, 2007, 88(1), 52–61.

[3] X. Chang, M. Liu, Z. Jin and J. Wang, *Studying on the impact of media coverage on the spread of covid-19 in hubei province, China*, Math. Biosci. Eng., 2020, 17(4), 3147–3159.

[4] T. Chen, J. Rui, Q. Wang et al., *A mathematical model for simulating the phase-based transmissibility of a novel coronavirus*, Infectious diseases of poverty, 2020, 9(1), 1–8.

[5] O. T. Chis, J. R. Banga and E. Balsa-Canto, *Structural identifiability of systems biology models: a critical comparison of methods*, PloS one, 2011, 6(11), e27755.

[6] O. Diekmann, J. A. P. Heesterbeek and J. A. Metz, *On the definition and the computation of the basic reproduction ratio r 0 in models for infectious diseases in heterogeneous populations*, Journal of mathematical biology, 1990, 28(4), 365–382.

[7] *Daily data on covid-19*. https://voice.baidu.com/act/newpneumonia/newpneumonia/?from=osari_aladin_banner#tab4.

[8] M. C. Eisenberg, S. L. Robertson and J. H. Tien, *Identifiability and estimation of multiple transmission pathways in cholera and waterborne disease*, Journal of theoretical biology, 2013, 324, 84–102.

[9] N. M. Ferguson, D. Laydon, G. Nedjati-Gilani et al., *Impact of non-pharmaceutical interventions (npis) to reduce Covid-19 mortality and healthcare demand*, 2020.

[10] M. Gatto, E. Bertuzzo, L. Mari et al., *Spread and dynamics of the Covid-19 epidemic in Italy: Effects of emergency containment measures*, Proceedings of the National Academy of Sciences, 2020, 117(19), 10484–10491.

[11] G. Giordano, F. Blanchini, R. Bruno et al., *Modelling the covid-19 epidemic and implementation of population-wide interventions in Italy*, Nature medicine, 2020, 26(6), 855–860.

[12] J. Goodman and J. Weare, *Ensemble samplers with affine invariance*, Communications in applied mathematics and computational science, 2010, 5(1), 65–80.

[13] J. Guedj, R. Thiébaut and D. Commenges, *Practical identifiability of HIV dynamics models*, Bulletin of mathematical biology, 2007, 69(8), 2493–2513.

[14] Z. He, L. Ren, J. Yang et al., *Seroprevalence and humoral immune durability of Anti-Sars-Cov-2 antibodies in wuhan, China: a longitudinal, population-level, cross-sectional study*, The Lancet, 2021, 397(10279), 1075–1084.

[15] J. P. La Salle, *The stability of dynamical systems*, SIAM, 1976.

[16] M. Li, G. Sun, J. Zhang et al., *Analysis of Covid-19 transmission in Shanxi province with discrete time imported cases*, Math. Biosci. Eng., 2020, 17(4), 3710.

[17] Q. Li, X. Guan, P. Wu et al., *Early transmission dynamics in Wuhan, china, of novel coronavirus–infected pneumonia*, New England journal of medicine, 2020.

[18] Q. Li, B. Tang, J. Wu et al., *Mathematical model reveals the influence of execution and adherence of mitigation strategies on the later period of Covid-19 and resumption of work*, Journal of Shaanxi Normal University (Natural Science Edition), 2020, 48(3), 1–6.

[19] P. Liu, S. He, L. Rong and S. Tang, *The effect of control measures on Covid-19 transmission in Italy: Comparison with guangdong province in china*, Infectious Diseases of Poverty, 2020, 9(1), 1–13.

[20] D. P. Lizarralde-Bejarano, D. Rojas-Díaz, S. Arboleda-Sánchez and M. E. Puerta-Yepes, *Sensitivity, uncertainty and identifiability analyses to define a dengue transmission model with real data of an endemic municipality of Colombia*, PloS one, 2020, 15(3), e0229668.

[21] H. Miao, C. Dykes, L. M. Demeter and H. Wu, *Differential equation modeling of HIV viral fitness experiments: model identification, model selection, and multimodel inference*, Biometrics, 2009, 65(1), 292–300.

[22] H. Miao, X. Xia, A. S. Perelson and H. Wu, *On identifiability of nonlinear ode models and applications in viral dynamics*, SIAM review, 2011, 53(1), 3–39.

[23] W. C. Roda, *Bayesian inference for dynamical systems*, Infectious Disease Modelling, 2020, 5, 221–232.

[24] W. C. Roda, M. B. Varughese, D. Han and M. Li, *Why is it difficult to accurately predict the Covid-19 epidemic?*, Infectious disease modelling, 2020, 5, 271–281.

[25] G. Schwarz, *Estimating the dimension of a model*, The annals of statistics, 1978, 461–464.

[26] N. Sugiura, *Further analysts of the data by Akaike's information criterion and the finite corrections: Further analysts of the data by Akaike's*, Communications in Statistics-theory and Methods, 1978, 7(1), 13–26.

[27] H. B. Taboe, K. V. Salako, J. M. Tison et al., *Predicting Covid-19 spread in the face of control measures in west Africa*, Mathematical biosciences, 2020, 328, 108431.

[28] B. Tang, X. Wang, Q. Li et al., *Estimation of the transmission risk of the 2019-ncov and its implication for public health interventions*, Journal of clinical medicine, 2020, 9(2), 462.

[29] N. Tuncer, H. Gulbudak, V. L. Cannataro and M. Martcheva, *Structural and practical identifiability issues of immuno-epidemiological vector–host models with application to rift valley fever*, Bulletin of mathematical biology, 2016, 78(9), 1796–1827.

[30] N. Tuncer and T. T. Le, *Structural and practical identifiability analysis of outbreak models*, Mathematical biosciences, 2018, 299, 1–18.

[31] N. Tuncer, C. Mohanakumar, S. Swanson and M. Martcheva, *Efficacy of control measures in the control of ebola, Liberia 2014–2015*, Journal of biological dynamics, 2018, 12(1), 913–937.

[32] *The average life expectancy of individual*. https://www.cia.gov.

[33] H. Wu, H. Miao, H. Xue et al., *Quantifying immune response to influenza virus infection via multivariate nonlinear ode models with partially observed state variables and time-varying parameters*, Statistics in Biosciences, 2015, 7(1), 147–166.

[34] Weikun, *A parallel implementation of mcmc*.

[35] W. Xia, T. Sanyi, C. Yong et al., *When will be the resumption of work in Wuhan and its surrounding areas during Covid-19 epidemic? A data-driven network modeling analysis*, Scientia Sinica Mathematica, 2020.