PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

An integrated framework for building trustworthy data-driven epidemiological models: Application to the COVID-19 outbreak in New York City

Sheng Zhang[1☯], **Joan Ponce**[1☯], **Zhen Zhang**[2☯], **Guang Lin**[1,3*], **George Karniadakis**[2]

1 Department of Mathematics, Purdue University, West Lafayette, Indiana, United States, **2** Division of Applied Mathematics and School of Engineering, Brown University, Providence, Rhode Island, United States, **3** School of Mechanical Engineering, Purdue University, West Lafayette, Indiana, United States

☯ These authors contributed equally to this work.
* guanglin@purdue.edu

Abstract

Epidemiological models can provide the dynamic evolution of a pandemic but they are based on many assumptions and parameters that have to be adjusted over the time the pandemic lasts. However, often the available data are not sufficient to identify the model parameters and hence infer the unobserved dynamics. Here, we develop a general framework for building a trustworthy data-driven epidemiological model, consisting of a workflow that integrates data acquisition and event timeline, model development, identifiability analysis, sensitivity analysis, model calibration, model robustness analysis, and projection with uncertainties in different scenarios. In particular, we apply this framework to propose a modified susceptible–exposed–infectious–recovered (SEIR) model, including new compartments and model vaccination in order to project the transmission dynamics of COVID-19 in New York City (NYC). We find that we can uniquely estimate the model parameters and accurately project the daily new infection cases, hospitalizations, and deaths, in agreement with the available data from NYC's government's website. In addition, we employ the calibrated data-driven model to study the effects of vaccination and timing of reopening indoor dining in NYC.

Author summary

The transmission dynamics of pandemics are often modeled by ordinary differential equations, which normally involve many undetermined parameters needed to be estimated from data. In this study, we provide a general framework, which includes identifiability analysis, sensitivity analysis, model robustness analysis, and uncertainty quantification, to examine the relationship between the model dynamics, data, and parameters. We apply our framework to the modeling of the COVID-19 outbreak in New York City and project the evolution of the pandemic.

OPEN ACCESS

Citation: Zhang S, Ponce J, Zhang Z, Lin G, Karniadakis G (2021) An integrated framework for building trustworthy data-driven epidemiological models: Application to the COVID-19 outbreak in New York City. PLoS Comput Biol 17(9): e1009334. https://doi.org/10.1371/journal.pcbi.1009334

Editor: Sergei L. Kosakovsky Pond, Temple University, UNITED STATES

Received: February 10, 2021

Accepted: August 10, 2021

Published: September 8, 2021

Data Availability Statement: The data used in this work can be found in the following public link: https://github.com/sheng-sheng-sheng/COVID.

Funding: SZ, JP, ZZ, GL, and GK gratefully acknowledge the support from ARO/MURI grant W911NF-15-1-0562. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. https://www.arl.army.mil/who-we-are/aro/.

Competing interests: The authors have declared that no competing interests exist.

Introduction

This study aims to answer a fundamental question: given epidemiological data, how to develop an appropriate model and identify which parameters we can accurately infer that would, in turn, allow us to correctly project the states of interest such as daily cases, hospitalizations, and deaths. The objective of this work is to provide a systematic way to model a pandemic accurately through carefully formulating a suitable model, uniquely identifying the model parameters, and projecting outbreaks under uncertainties based on the different epidemiological data available. To address the above fundamental question, we propose a general integrated framework to approach the problem systematically through identifiability analysis, sensitivity analysis, model robustness analysis, and projection under uncertainties.

Numerous modeling approaches have been used to gain insight into epidemic disease's ever-evolving dynamics and the effects the interventions have had on containing the spread. Mathematical modeling is an efficient way to test and evaluate the effectiveness of hypothetical interventions that cannot be tested out due to practical or ethical limitations. Describing the disease's development involves representing a highly complex process affected by social, environmental, and biological factors. Compartmental models have traditionally been used to depict systems that include individuals with different health statuses that change in time. In particular, mathematical modeling of COVID-19 using compartmental models described by ordinary differential equations (ODEs) such as susceptible–infectious–recovered (SIR) [1–3], modified SIR [1, 4–6], susceptible–exposed–infectious–recovered (SEIR) [7–9] and modified SEIR models [10, 11] has been used extensively in an attempt to capture the virus' spread. These types of lumped mechanistic models, unlike data-driven models, can explore future outcomes of the pandemic and evaluate the effects of various interventions. Compartmental models are commonly applied in epidemiology as they are simple and easily tractable. However, their accuracy is constrained by parameter uncertainties and gaps in information about the disease dynamics. Likewise, assumptions to maintain model simplicity may affect the estimated values. For a long-lasting pandemic, the model parameters change with time; hence the parameter identification problem becomes nontrivial given the fact that typically a limited amount of relevant data is available.

In an ODE-based epidemiological model, the system parameters usually contain critical information that often cannot be measured directly, such as the transmission rate, which needs to be inferred from data. A necessary condition for the well-posedness of a parameter estimation problem in ODE theory is the model's structural identifiability if we assume noise-free data. The structural identifiability analysis can be performed without any experimental data; it addresses whether the parameter estimation problem is well-posed under ideal conditions. Should the postulated model not be structurally identifiable, the parameters obtained will be unreliable. However, a model can be structurally identifiable (a necessary condition) but may not be practically identifiable. Thus, the structural identifiability analysis may conclude that a model's parameters are uniquely determined, yet when real-life, noisy data are used, the estimated parameter values could still be unreliable. To conduct the practical identifiability analysis, we compute the correlation matrices of model parameters in different settings using Fisher Information Matrices (FIMs) following lines of approach in [12, 13].

Non-identifiability is a problem frequently encountered in pandemics modeling since, typically, not every state variable is available. In recent literature, model identifiability issues have been studied due to the wide variation in model projections in the context of the COVID-19 pandemic [2, 14–16]. Tuncer et al. analyzed the structural and practical identifiability of some of the most widely-used pandemic models, including SIR, SIR with treatment, and SEIR, assuming only one observed data type is available using simulated data [17]. Roda et al.

extended these ideas by studying SIR and SEIR models' practical identifiability using data from the COVID-19 outbreak in Wuhan, using only the counts of infected individuals as the available data [2]. They found that complex mechanistic models are more likely to have identifiability issues compared to simpler models. Massoni et al. provided a systematical structural identifiability and observability analysis of 255 available compartmental models for COVID-19 and found that approximately one-third of them have structurally non-identifiable parameters [14]. Therefore, an identifiability test should be conducted as a sanity check when a new ODE-based mechanistic model is proposed, to ensure trustworthy parameter estimation.

Furthermore, precise estimates of parameters allow us to determine an epidemiologically relevant value called the basic reproduction number, $\mathcal{R}_0$. The basic reproduction number is defined as the average number of infections caused by one infected individual in an entirely susceptible population when disease control is absent, and it determines whether an outbreak will occur.

Still, uncertainty about parameter values could be relatively high at the beginning of an outbreak even if identifiability is guaranteed. Therefore, determining the response of a model's output to parameter variation helps identify sources of uncertainty. Sensitivity analysis studies how the uncertainty in the model's output can be allocated to different inputs' uncertainty sources [18]. It allows a better understanding of the model to analyze how the model parameters affect the output.

In the present work, we propose to integrate these steps of identifiability and sensitivity analysis together with policy change timelines, data availability, and uncertainties in projection. We apply the proposed general framework to introduce a modified SEIR model, which we extend to include vaccination, and project the transmission dynamics of COVID-19 in New York City (NYC) under vaccination and different safety measures relaxation scenarios. Daily cases, hospitalizations, and deaths in NYC are used to demonstrate the way to employ the proposed framework for simulating the ongoing COVID-19 pandemic in the city from early 2020 until February 2021.

COVID-19, which emerged in China in late 2019, has caused an outbreak affecting over 200 countries worldwide and was declared a pandemic on March 11, 2020 [19] by the World Health Organization. The strategies to control the spread of the virus in 2020 (before a vaccine was available) were mainly directed towards non-pharmaceutical interventions, such as isolation of infected individuals, social distancing, and face-mask use. COVID-19 was detected in a patient in NYC in early 2020 [20, 21]. The high population density and lack of control measures in the first three weeks produced an exponential increase in cases, which exceeded 20,000 before the statewide stay-at-home order was put in place on March 22, 2020 [22]. Between March 22, 2020, and June 8, 2020, a host of social distancing measures and non-essential businesses closings in NYC lowered the incidence from almost 2,000 daily cases in April to a couple hundred daily cases by June 8, 2020 [23]. The city's first phase of its four-phase reopening plan began on June 8, 2020, and the final stage started on July 20, 2020.

The proposed model considers presymptomatic, asymptomatic, hospitalized, isolated, and deceased individuals. Structural identifiability, practical identifiability, and sensitivity of the model are studied. Once the parameters that can be reliably estimated are identified, the parameter estimation portion of the study is broached. Then, the NYC outbreak data (daily infected, hospitalized, and deceased individuals) are employed to estimate the model parameters to understand how public policy such as isolation, public closings, and other social distancing measures impact the transmission dynamics. Moreover, confidence intervals are computed, and the model's projective capabilities are analyzed by considering the uncertainty of policies in constant flux. Next, we further investigate model robustness. The main modeling assumption is that the transmission rate is changing in time, following different policy measures, in a piecewise fashion. The vaccination deployment in NYC is incorporated into the model, and we evaluate the combined effect of vaccination and the gradual reopening process currently underway.

The novelty of this work is multi-fold:

- We develop a general framework and workflow for building a trustworthy data-driven epidemiological model;
- We propose a modified SEIR model with vaccination and validate it with the pandemic data in New York City (daily cases, hospitalizations, and deaths); every parameter in our model has physical meaning;
- We systematically study the structural identifiability, practical identifiability, and sensitivity to examine the relationship between the model dynamics, data, and parameters;
- We treat the transmission rate, hospitalization ratio, and death from hospital ratio as time-dependent model parameters based on the event-timeline, and calibrate the identifiable model parameters using simulated annealing and MCMC simulations. We also investigate model robustness by studying how the model behaves under random perturbations;
- We demonstrate the model's projective capabilities under uncertainties for different future scenarios;
- We specifically investigate the effects of indoor dining reopening and vaccination scenarios as a reference for policymakers.

General framework and workflow

Holmdahl and Buckee, in [24], discussed different types of models for the COVID-19 epidemic as well as the distinct challenges in these approaches. The authors highlighted that "models are a way to formalize what we know about the viral transmission and explore possible futures of a system that involves nonlinear interactions, something that is almost impossible to do using intuition alone." They further elaborate that "models will be useful for exploring possibilities rather than making strong projections about longer-term disease dynamics." Thus, a systematic way of designing an effective data-driven model is essential for assessing ongoing control strategies endowed with uncertainty quantification.

To systematically design an effective data-driven model, we propose a general framework for building a trustworthy data-driven epidemiological model, which constructs a workflow to integrate data acquisition and event timeline, model development, identifiability analysis, sensitivity analysis, model calibration, model robustness analysis, model projection with uncertainties and investigation of reopening scenarios together. Fig 1 gives an overview of the framework, with details for each stage provided below.

(I). Acquire data and look for major interventions or events that could affect the transmission dynamics of the epidemic.

(II). Develop an epidemiological model that accommodates the data and events in (I). The intervention events can be encoded through time-dependent parameters in the model.

(III). Use both structural and practical identifiability analysis to determine the parameters to fit. Domain knowledge can also be incorporated in this step to help choose the parameters. If the model is not identifiable but one prefers unique parameters, one should fix some of the non-identifiable parameters in the model or propose other models. Otherwise, one can proceed to (IV).

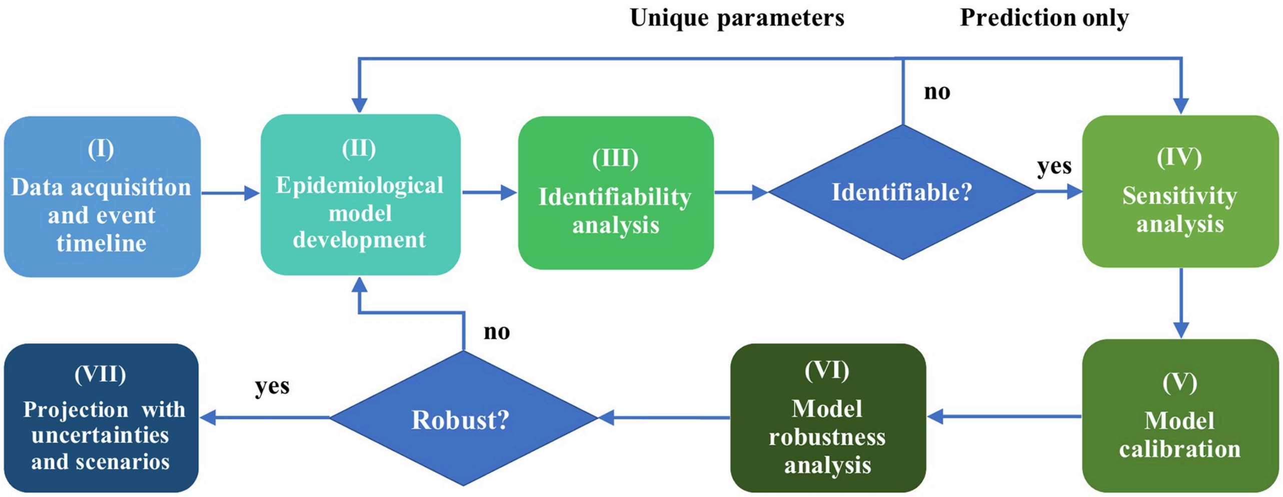
Unique parameters
Prediction only
no
(I) Data acquisition and event timeline
(II) Epidemiological model development
(III) Identifiability analysis
Identifiable?
yes
(IV) Sensitivity analysis
no
(V) Model calibration
(VI) Model robustness analysis
Robust?
yes
(VII) Projection with uncertainties and scenarios

Fig 1. A general framework for building a trustworthy data-driven epidemiological model—An overview of the main contribution. In this work, we propose a general framework for building a trustworthy data-driven epidemiological model, which constructs a workflow to integrate data acquisition and event timeline, model development, identifiability analysis, sensitivity analysis, model calibration, model robustness analysis, and projection with uncertainties and scenarios. We first introduce a modified SEIR model that accommodates the pandemic data in New York City. Secondly, we study the structural identifiability, practical identifiability, and sensitivity to examine the relationship between the model's data and parameters. We then calibrate the identifiable model parameters using simulated annealing and MCMC simulation. Model robustness is then checked to study how the model behaves under random perturbations. In addition, we demonstrate the model's projective capabilities with uncertainties. Finally, reopening scenarios are investigated as a reference for policymakers.

https://doi.org/10.1371/journal.pcbi.1009334.g001

(IV). Conduct sensitivity analysis to find the most sensitive parameters to each observable. In cases when an observable is insensitive to a parameter, even if the parameter is non-identifiable, one can still use the model to calibrate that observable.

(V). Calibrate the most sensitive and identifiable model parameters. Estimate the model compartments and the reproduction number.

(VI). Check model robustness assuming different types of noise in the data. If the model is robust to noise, then one can proceed to (VII), otherwise one should go back to (II) and fix some model parameters or employ other models.

(VII). Project the future development of the epidemic with uncertainties assuming the current control measures. Then investigate how policy changes could influence the transmission dynamics of the epidemic.

We evaluate the effectiveness of the proposed framework by applying all the outlined steps to the outbreak dataset in NYC. One of the advantages of the framework's generality is that it is not limited to a single dataset or model. In general, it provides a guideline on how to build an effective and trustworthy epidemiological model with the available data. For illustration purposes, we show how our framework can handle different data types by assuming scenarios when some observables in the NYC dataset are missing. However, for practical use, one should determine the data that are fed to the model at the very beginning.

(I) Data acquisition and event timeline

We use the data consisting of daily cases, hospitalizations, and deaths between February 29, 2020, and February 4, 2021, to fit the model's parameters. All the data used in this paper were extracted from the NYC's government's website and collected by the NYC Health Department

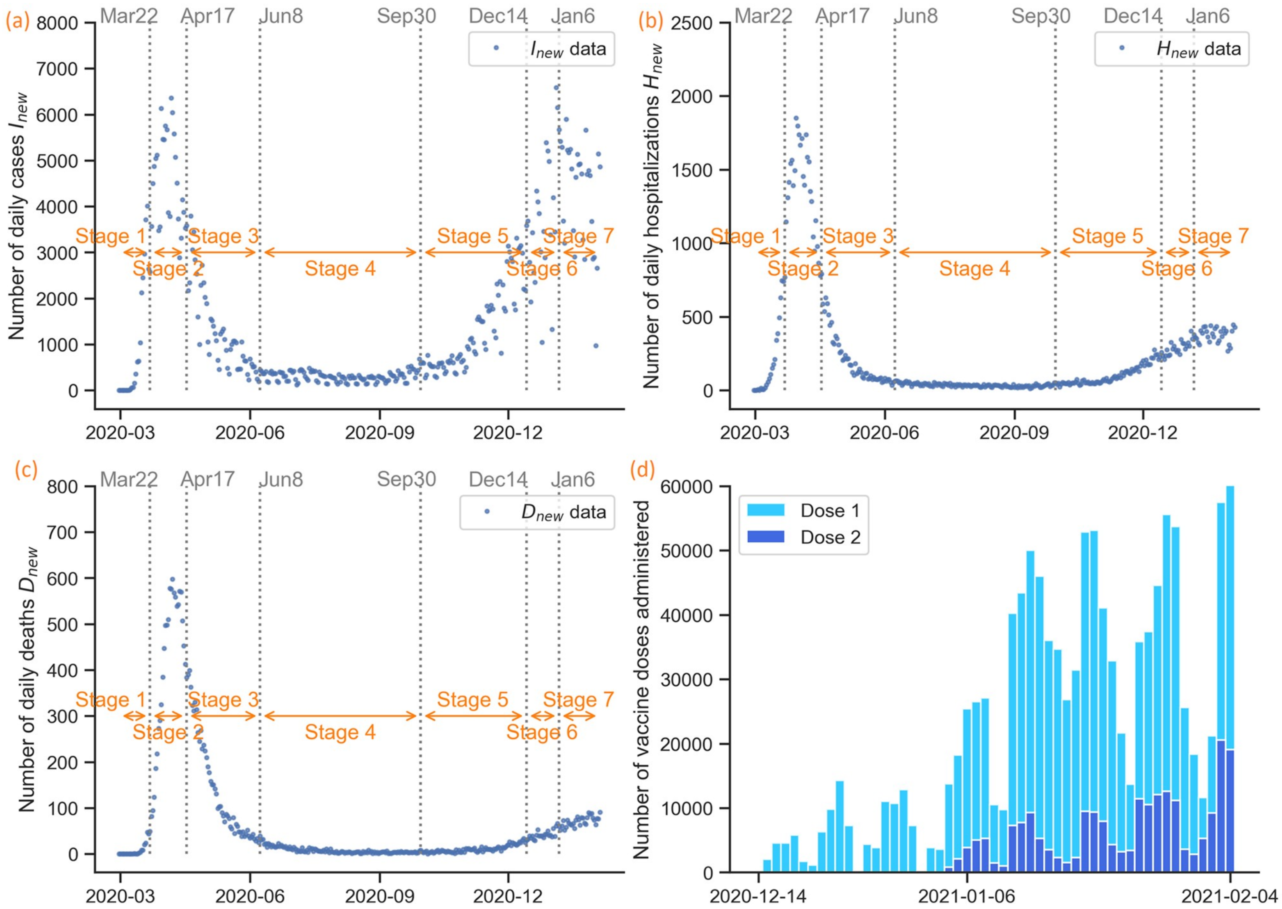

Fig 2. COVID-19 epidemic in New York City: Data and event timeline. (a) Daily confirmed cases (February 29, 2020–February 4, 2021). A person is classified as a confirmed COVID-19 case when they test positive in a molecular test (PCR). We split the data into seven time periods based on interventions implemented. The starting times of interventions are shown on the top of each subfigure. (b) Daily hospitalized population (February 29, 2020–February 4, 2021). (c) Daily deceased population. (February 29, 2020–February 4, 2021). A deceased individual is classified as a disease-related death if they had a positive PCR test for the virus within the last 60 days. (d) Daily vaccinated population. (December 14, 2020–February 4, 2021).

https://doi.org/10.1371/journal.pcbi.1009334.g002

[23]. Hospitalization data were collected from several sources, such as NYC public hospitals, non-public hospitals, and the Health Department's syndromic surveillance database, which track hospital admissions across NYC. The data on the NYC online repository [23] are published by the date of the event, rather than the date of the report. A person is classified as a confirmed COVID-19 case when they test positive in a molecular test (PCR). A deceased individual is classified as a disease-related death if they had a positive PCR test for the virus within the last 60 days.

Since the first case of COVID-19 was reported in NYC on February 29, 2020, local authorities have implemented several non-pharmaceutical interventions, such as social distancing and mask-wearing [22, 23]. These restrictions were later relaxed as the incidence decreased. Any control measures implemented would have an impact on the transmission term β. Thus, we identified seven time periods defined based on the interventions implemented (also see Fig 2):

- **Stage 1 (no control): February 29, 2020–March 22, 2020**
  The New York State governor declared a state of emergency on March 7, 2020, after 89 positive cases were identified [25]. However, most businesses operated as usual until March 14,

2020, when some public libraries closed. Nightclubs, theaters, and concert venues followed suit on March 17, 2020.

- **Stage 2 (stay-at-home order): March 22, 2020–April 17, 2020**
  A stay-at-home order (also known as PAUSE) issued by the New York State's governor's office went into effect on March 22, 2020. The PAUSE plan comprised a 10-point policy that mandated all non-essential businesses statewide to close, cancel gatherings of any size, and businesses that provide essential services must facilitate social distancing of at least six feet, among others. Schools and universities closed and moved to remote instruction.

- **Stage 3 (mask mandate): April 17, 2020–June 8, 2020**
  New York State was one of the first states to issue orders mandating face coverings in public spaces [26]. This decision followed the Centers for Disease Control and Prevention (CDC) guidelines, which encouraged people to wear masks to prevent transmission of the virus through droplets generated when an infected person coughs or sneezes [27].

- **Stage 4 (four-phase reopening): June 8, 2020–September 30, 2020**
  The stay-at-home order was effective in bringing down the disease's incidence. As a result, a four-phase reopening plan was developed, taking into account seven health metrics the city needed to meet before reopening [28]. NYC entered Phase 1 of reopening on June 8, Phase 2 on June 22, Phase 3 on July 6, and Phase 4 on July 20. Each phase had specific policies that determined what businesses could reopen and in what capacity. Industries that posed the lowest risk of infection for employees and customers were allowed to reopen in Phase 1. These included but were not limited to construction, manufacturing, and wholesale supply-chain businesses and retailers for curbside pickup, in-store pickup, or drop-off [29]. Phase 2 allowed offices, places of worship (25% capacity), finance and insurance, administrative support, among others, to reopen if they follow established social distancing guidelines [30]. In Phase 3, some personal services were allowed to reopen; indoor dining was not allowed in NYC, even though this phase allowed other areas of the state to enable indoor dining at a reduced capacity. NYC entered reopening Phase 4 on July 20. Restrictions regarding group gatherings were eased, and meetings of up to 50 people were allowed. Indoor religious meetings were allowed to resume at 33% capacity. Malls, zoos, and botanical gardens were also permitted to reopen in this phase. NYC stayed at Phase 4 until September 30, 2020, when indoor dining at 25% capacity was allowed.

- **Stage 5 (indoor dining reopens at reduced capacity): September 30, 2020–December 14, 2020**
  The careful reopening process, which followed the strict stay-at-home order, maintained a low infection rate in the city (below 1%) [31]. NYC resumed indoor dining services at 25% capacity on September 30, 2020, intending to double the capacity if infection rates remained low. Restaurants were required to follow an extensive set of rules upon reopening, such as temperature checks, contact tracing reporting, mask usage except when seated, and a midnight curfew.

- **Stage 6 (indoor dining closes and vaccination begins): December 14, 2020–January 6, 2021**
  In December 2020, the increasing rate of virus transmission in NYC threatened to overwhelm hospital capacity. Although contact tracing data from NYC placed indoor dining as the fifth source of new infections in the state, the CDC designated indoor dining as a "high risk" activity [32]. The governor's decision to ban indoor dining was an attempt to halt the steep increase in cases and avoid a broader shutdown. In the same week when the

governor's office closed indoor dining again, the first coronavirus vaccine was administered in Queens on December 14, 2020 [33].

- **Stage 7 (Christmas and New Year holiday ends): January 6, 2021–February 14, 2021** The increased social activities during Christmas and New Year celebrations had a significant impact on the outbreak's spread. This final period is defined by an event, the end of the holidays, and restaurants' reopening in limited capacity announced for February 14, 2021.
  Two shipments from drug companies Pfizer and Moderna aim to cover a quarter of the estimated 1.8 million people deemed high priority to receive the vaccine in the first phase of distribution in the state. However, even though vaccination started on December 14, 2020, people vaccinated do not develop immunity to the virus immediately. The Pfizer vaccine's first dose needs about 14 days to be 52% effective. The second dose should be administered three weeks after the first dose. The reported effectiveness of the Pfizer vaccine is 95%, while Moderna reports 94.1% when two doses are received [34, 35]. It is worth noting that if a person only receives one dose, its effectiveness varies depending on the company that produced the vaccine. For example, the Pfizer-BioNTech vaccine is roughly 52% effective after the first dose, while the Moderna vaccine can provide 80.2% protection after one dose [34]. At–risk groups, such as older people in nursing homes and public health professionals, are prioritized for vaccination in NYC in this first phase. We use data reported by NYC Health, extracted from the Citywide Immunization Registry (CIR). The data include the daily numbers of individuals who have received the first and second dose [33]; see Fig 2.

(II) Epidemiological model development

Classic SIR and SEIR models do not include the exposed and presymptomatic periods, which play an important role in this particular disease. Recent studies have revealed the role of asymptomatic [39–42] and presymptomatic individuals [43–45] in the disease's transmission chain. Additionally, due to CDC recommendations, infected individuals are required to self-isolate once they test positive for ten days [27]. We put these individuals in the isolated compartment of our model. Therefore, to account for the complete epidemiological characteristics of a COVID-19 infection, in this model we modify an SEIR model to include presymptomatic, asymptomatic, hospitalized, isolated, and deceased compartments; see Fig 3.

These modifications allow a more accurate description of the biology of the disease. Since this study focuses on a single outbreak, births and other deaths are not considered. When the epidemic begins, all individuals are susceptible and transit to the exposed class via contact with presymptomatic, symptomatic, or asymptomatic individuals. Moreover, we assume that there is no reinfection. In other words, once an individual recovers, they do not become susceptible again. In addition, we do not include disease transmission from hospitalized individuals. We divide the total population into nine different compartments: susceptible (S), exposed (E), presymptomatic infected (P), symptomatic infected (I), asymptomatic infected (A), hospitalized (H), isolated (Q), deceased (D), and recovered (R). The contribution to the transmission of COVID-19 from asymptomatic individuals (A) relative to the transmission from symptomatic individuals (I) is labeled as ϵ. Therefore, if $\epsilon = 0.75$, this means that an asymptomatic individual is 75% as infectious as a symptomatic individual. Similar to [37], we assume that presymptomatic individuals (P) are just as infectious as asymptomatic individuals (A). The transmission rate of the disease is denoted by β. After a latent period ($1/d_E$), an exposed individual becomes

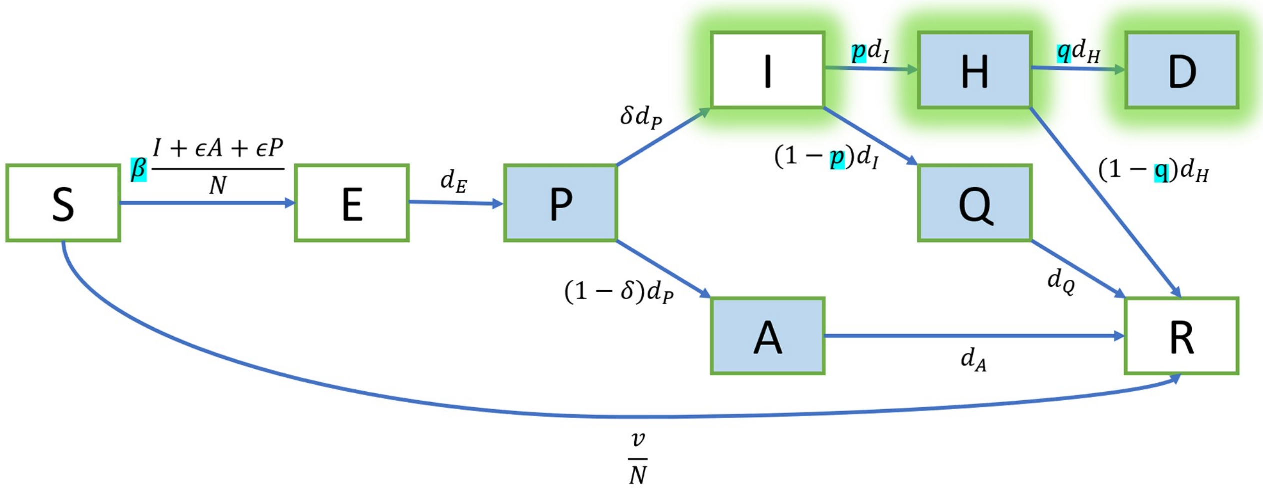

Fig 3. Transition diagram between epidemiological classes. We modify the classic SEIR model to include presymptomatic (P), asymptomatic (A), hospitalized (H), isolated (Q), and deceased (D) individuals. The given data are the inflows of symptomatic (I), hospitalized (H), and deceased (D) individuals. The parameters to estimate are (β, p, q). See Table 1 for the notations and the initial values. See Table 2 for the parameters. See Eq (1) for the corresponding ODE system.

https://doi.org/10.1371/journal.pcbi.1009334.g003

presymptomatic infectious. A presymptomatic infectious individual develops symptoms with probability δ or is asymptomatic with probability $(1 - \delta)$ after $(1/d_P)$ days. Asymptomatic individuals recover after an infective period of $(1/d_A)$. A proportion p of symptomatic individuals are hospitalized after an infective period of $(1/d_I)$, while the rest of them are isolated (for example, isolated at home) but do not go to the hospital. The isolation period is $(1/d_Q)$. The hospitalization period is $(1/d_H)$, at the end of which, a proportion q of the hospitalized individuals die while the rest of them recover. Asymptomatic cases pose a challenge to identify since there is no widespread systematic testing of the population [39]. Thus, there is a wide range of estimates for the proportion of symptomatic individuals δ. In this study, we use the best estimate of δ provided by the CDC [36].

Moreover, the proposed model has considered the vaccination deployment in NYC and the impact on the city's daily cases, hospitalizations, and deaths. A susceptible individual

Table 1. Notations and initial values for the model in Fig 3.

Notation	Meaning	Initial value	Note
N	Total population	8399000	We assume constant total population in NYC
S	Susceptible	8398713	$S(0) = N - E(0) - P(0) - I(0) - A(0) - H(0) - Q(0) - D(0) - R(0)$
E	Exposed	270	$E(0) = E_*/\delta$, where E_* is the number of cases from day $(2/d_P + 1)$ to day $(2/d_P + 2/d_E)$ (Mar 6–10, 2020)
P	Presymptomatic	15	$P(0) = P_*/\delta$, where P_* is the number of cases from day 1 to day $2/d_P$ (Mar 1–5, 2020)
I	Symptomatic	1	$I(0) = 1$ from data
A	Asymptomatic	1	$A(0) = I(0) \times (1 - \delta)/\delta \approx 1$
H	Hospitalized	0	$H(0) = 0$ from data
Q	Isolated	0	$Q(0) = 0$
D	Deceased	0	$D(0) = 0$ from data
R	Recovered	0	$R(0) = 0$
I_{sum}	Cumulative cases	1	$I_{sum}(0) = 1$ from data
H_{sum}	Cumulative hospitalizations	0	$H_{sum}(0) = 0$ from data
D_{sum}	Cumulative deaths	0	$D_{sum}(0) = 0$ from data

https://doi.org/10.1371/journal.pcbi.1009334.t001

Table 2. Parameters for the model in Fig 3.

Parameter	Meaning	Calibration	Range/Value	Note
β	Transmission rate	fitted	[0, 1]	Fitted within wide range
p	Hospitalization ratio	fitted	[0, 1]	By definition
q	Death from hospital ratio	fitted	[0, 1]	By definition
ϵ	Infectivity ratio of asymptomatic to symptomatic	fixed	0.75	Ref. [36]
δ	Proportion of symptomatic individuals	fixed	0.6	Ref. [36]
$1/d_E$	Latent period	fixed	2.9 days	Ref. [37]
$1/d_P$	Mean infectious period of P class	fixed	2.3 days	Ref. [37]
$1/d_I$	Mean infectious period of I class	fixed	2.9 days	Ref. [37]
$1/d_A$	Mean infectious period of A class	fixed	7 days	Ref. [38]
$1/d_H$	Mean duration of H class	fixed	6.9 days	Ref. [36]
$1/d_Q$	Mean duration of Q class	fixed	10 days	Ref. [27]

https://doi.org/10.1371/journal.pcbi.1009334.t002

leaves the (S) class and joins the recovered individuals (R) once they are effectively vaccinated, i.e., they are vaccinated and the vaccine prevents them from getting the disease in the future. To calculate the number of effectively vaccinated individuals that takes into account the vaccine efficacy, we use a weighted sum of the number of the first doses and second doses administered. See Fig 3, where v is defined as the daily number of effectively vaccinated individuals.

The ODE system for our model (see Fig 3) is the following:

$$\begin{cases} \frac{dS}{dt} &= -\beta \frac{I+\epsilon A+\epsilon P}{N} S - \frac{v}{N} S \\ \frac{dE}{dt} &= \beta \frac{I+\epsilon A+\epsilon P}{N} S - d_E E \\ \frac{dP}{dt} &= d_E E - d_P P \\ \frac{dI}{dt} &= \delta d_P P - d_I I \\ \frac{dA}{dt} &= (1-\delta) d_P P - d_A A \\ \frac{dH}{dt} &= p d_I I - d_H H \\ \frac{dQ}{dt} &= (1-p) d_I I - d_Q Q \\ \frac{dD}{dt} &= q d_H H \\ \frac{dR}{dt} &= d_A A + (1-q) d_H H + d_Q Q + \frac{v}{N} S. \end{cases} \tag{1}$$

Note that the compartments (I, H, D) represent the current symptomatic individuals, hospitalizations, and deaths. In order to represent the daily cases I_{new}, hospitalizations H_{new}, and deaths D_{new}, we add the following ODEs that take into account the inflow of the these compartments to record the cumulative cases I_{sum}, hospitalizations H_{sum}, and deaths D_{sum}:

$$\begin{cases} \dfrac{dI_{sum}}{dt} & = \delta d_P P \\ \dfrac{dH_{sum}}{dt} & = p d_I I \\ \dfrac{dD_{sum}}{dt} & = q d_H H. \end{cases} \tag{2}$$

Now, the daily numbers are just the increments of the cumulative numbers:

$$\begin{cases} I_{new}(t) & = I_{sum}(t) - I_{sum}(t-1) \\ H_{new}(t) & = H_{sum}(t) - H_{sum}(t-1) \\ D_{new}(t) & = D_{sum}(t) - D_{sum}(t-1) \end{cases} \tag{3}$$

for $t = 1, 2, 3, \cdots$.

Time-dependent model parameters (piecewise constant *β*, *p*, and *q*)

The transmission rate of a disease, β, is the per capita rate of infection when a contact occurs. Directly measuring the transmission rate is not possible for most infections [46]. Nevertheless, if we want to quantify the effects of public health policies that directly impact the transmission rate, estimating this value accurately is critical. Moreover, public health policy and the discovery of better therapies and treatments affect other parameters besides the disease's transmission rate. Notably, the percentage of disease-related deaths changes over the course of the outbreak [47]. Similarly, the hospitalization ratio varies due to increased resources channeled to the healthcare system in the city [23, 48]. Control measures implemented in NYC and the subsequent relaxation of restrictions impact the incidence curve in different ways—most of them non-linear. Therefore, defining the transmission rate $\beta(t)$ as a piecewise constant function is a simplification that allows us to estimate the impact each policy has in each stage. Similarly, we define the piecewise constant hospitalization ratio $p(t)$ and death from hospital ratio $q(t)$, which also exhibit varying values over time:

$$\beta(t), p(t), q(t) = \begin{cases} \beta_1, p_1, q_1 & t \in \text{Stage } 1 \\ \beta_2, p_2, q_2 & t \in \text{Stage } 2 \\ \beta_3, p_3, q_3 & t \in \text{Stage } 3 \\ \beta_4, p_4, q_4 & t \in \text{Stage } 4 \\ \beta_5, p_5, q_5 & t \in \text{Stage } 5 \\ \beta_6, p_6, q_6 & t \in \text{Stage } 6 \\ \beta_7, p_7, q_7 & t \in \text{Stage } 7. \end{cases} \tag{4}$$

We fit the parameters in each stage defined by policy changes, such as the stay-at-home order and the subsequent reopening processes; see Fig 2.

Reproduction number

The basic (control) reproduction number, denoted by $\mathcal{R}_0$ ($\mathcal{R}_c$), is the average number of secondary infections caused by one infected individual in an entirely susceptible well-mixed population in the absence (presence) of disease control. The control reproduction number of the model is:

$$\mathcal{R}_c = \beta\left[\frac{\epsilon}{d_P} + \frac{\delta}{d_I} + \frac{(1-\delta)\epsilon}{d_A}\right]. \tag{5}$$

In this particular study, given that the model parameters are defined in a piecewise fashion and there was no control in Stage 1, the basic reproduction number, $\mathcal{R}_0$, is computed by using $\beta = \beta_1$.

The transmission of the disease slows down when there are more immune individuals. Since $\mathcal{R}_c$ is the number in an entirely susceptible population, we can calculate the effective reproduction number:

$$\mathcal{R}_e = \mathcal{R}_c \cdot \frac{S}{N}. \tag{6}$$

By setting $\mathcal{R}_e = 1$, we obtain the immunity threshold of the ODE system, which is the critical portion of the population needed to be immune to stop the transmission of the disease:

$$\mathrm{IT} = 1 - \frac{1}{\mathcal{R}_c}. \tag{7}$$

The herd immunity threshold (HIT) is calculated by substituting $\mathcal{R}_c$ with $\mathcal{R}_0$. A higher $\mathcal{R}_0$ results in a higher HIT.

(III) Identifiability analysis

In this section we address whether a set of unknown parameters in the proposed model is globally identifiable from the available data. Fitting a model to the data is not sufficient to show how reliable the estimated parameters are. Insufficient or noisy data can produce drastically different sets of parameters without affecting the fit to data if a model is non-identifiable [2]. Furthermore, depending on the available data (observables), different models may be appropriate.

Formally speaking, a parameter in a dynamical system is considered to be identifiable if the solutions can uniquely determine it. Two different types of identifiability, namely structural and practical identifiability are considered in this paper. Structural identifiability analysis studies the uniqueness of parameter values from the perspective of the structure of the equations and is normally conducted before the fitting of the model, thus commonly referred to as a priori identifiability. Global (structural) identifiability provides conclusions about a parameter's identifiability in the entire parameter space [49–52]. In particular, it guarantees the opportunity of uniquely identifying the model parameters from the data [49–51, 53]. In some cases, however, local structural identifiability may be sufficient, and hence the range of values of the parameter to be identified should be limited. On the other hand, practical identifiability analysis mainly addresses the issue of nonuniqueness when fitting the model on the discrete data points, i.e., a posteriori. Structural identifiability does not imply practical identifiability because of the amount and quality of the data. A detailed explanation of these two types of identifiability can be found in S4 Text. We use the open-source software SIAN [54] and GenSSI2.0 [55] for structural identifiability and use correlation matrix calculated from Fisher Information Matrix (FIM) for practical identifiability. Details of the implementation can also be found in the Supporting information.

In our framework, we analyze both structural and practical identifiability, and use the results as guidelines for parameter selection. The importance of performing both types of analysis resides in the fact that structural identifiability itself does not guarantee the goodness-of-fit of the model. It turns out that in our case, fitting all structural identifiable parameters would lead to practically non-identifiable results.

Structural identifiability

There are 11 undetermined parameters in the proposed model, and it is impossible to fit every parameter without fixing some of the values. For example, it is unnecessary to fit biologically determined parameters such as the time an individual spends in the exposed or infected classes. Since d_E, d_P, d_I, d_A, d_H, and d_Q are determined by the biology of the disease, we fix these values according to [37]. The initial conditions of all state variables are fixed as in Table 1.

We analyze the structural identifiability of the rest of the parameters when different types of data are given. Note that for general use of our modeling framework, one should fix the dataset at Step (I). Here we consider all the scenarios just for illustration purpose. Specifically, we assume that the data are given as the cumulative cases I_{sum}, cumulative hospitalizations H_{sum}, and cumulative deaths D_{sum}, or a subset of the three aforementioned observables, because these quantities can be calculated directly from daily quantities I_{new}, H_{new}, and D_{new}. In other words, it is equivalent to assume X_{new} or X_{sum} to be given as one of the observables, where X can be I, H, and D. The effectively vaccinated population v is treated as an input variable to the system. According to Table 3, when H_{sum} or D_{sum} is not available, the model is not identifiable and the fitting result will not be unique. The results are to be interpreted in the following way. In the case of lacking deceased individual counts, the death from hospital ratio q cannot be inferred accurately. If the hospitalization data are not available, neither the hospitalization ratio p nor the death from hospital ratio q may be inferred accurately.

The analysis above shows that it is hard to draw conclusions about the fitting correctness of transmission rate, proportion of isolated individuals, and proportion of disease-related deaths when H_{sum} or D_{sum} is missing. One of the main differences between the proposed model and most other existing SEIR-based models is that our model integrates information of infectious, hospitalized, and deceased populations simultaneously, therefore producing more reliable results on these estimated parameter values. Since we have data for all three observables in NYC, we should utilize all of them.

In practice, hospitalization data could be reported in different ways; some databases provide daily reports of the number of hospitalized individuals, whereas others register the number of currently hospitalized individuals. Regardless of the data type available, structurally identifiability of the model remains the same according to S2 Table.

Practical identifiability

We then proceed with fitting 5 undetermined parameters using all the available data, i.e., I_{sum}, H_{sum}, and D_{sum}. The model-fitting techniques, including the loss function and optimization method, are detailed in the next section. The fitted parameter values can be found in Table 4.

Table 3. Structural identifiability of β, p, q, ϵ, δ with different observables. Global/not means structurally globally/not identifiable, respectively. We fix all the rest of the parameters as in Table 2. We fix the initial condition of each state variable as in Table 1.

Parameter	I_{sum}, H_{sum}, D_{sum}	I_{sum}, H_{sum}	I_{sum}, D_{sum}	H_{sum}, D_{sum}	I_{sum}	H_{sum}	D_{sum}
β	global	global	global	global	global	global	global
p	global	global	not	global	not	global	not
q	global	not	not	global	not	not	not
ϵ	global	global	global	global	global	global	global
δ	global	global	global	global	global	global	global

https://doi.org/10.1371/journal.pcbi.1009334.t003

Table 4. Practical identifiability and estimation of parameters when fixing d_E, d_P, d_I, d_A, d_H, d_Q. The symbol ✓/✗ means practically identifiable/not identifiable, respectively. The fitted values will **not** be counted towards our final result because the model is not identifiable in this case.

Parameter	Identifiable	Stage 1	Stage 2	Stage 3	Stage 4	Stage 5	Stage 6	Stage 7
β	✗	0.64	0.25	0.32	0.19	0.59	0.20	0.43
p	✓	0.38	0.31	0.15	0.12	0.08	0.08	0.08
q	✓	0.14	0.38	0.35	0.22	0.13	0.16	0.19
ϵ	✗	1.00	0.42	0.05	0.76	0.00	1.00	0.00
δ	✗	1.00	0.95	0.68	0.62	0.77	0.64	0.78

https://doi.org/10.1371/journal.pcbi.1009334.t004

We see that the values of ϵ and δ vary a lot among different stages, which is inconsistent with the reality. This poses a question on the practical identifiability of the model.

There are two approaches commonly applied to determine the practical identifiability of ODE models, namely Monte Carlo methods and Fisher information matrix (FIM) based methods, with details given in S4 Text. Monte Carlo methods are computationally heavy and could produce unreliable results when the number of undetermined parameters is large. On the other hand, it is easier to use FIM computationally, even in high dimensions. Thus, we suggest to apply FIM-based methods first to determine a set of parameters that are not practically identifiable and fix them (or only use the identifiable combinations). Once these parameters are fixed, one can then apply the Monte Carlo methods to check whether the rest of the parameters are identifiable or not. To distinguish those two approaches, we refer to Monte Carlo methods for determining practical identifiability as model robustness analysis, which is detailed in (VI). The calculation of FIM and the correlation matrices is given in S4 Text.

As shown in Fig 4(b), there is a strong correlation between ϵ, δ, and β, while either p or q is uncorrelated with the rest of the parameters. The same phenomenon is observed when we

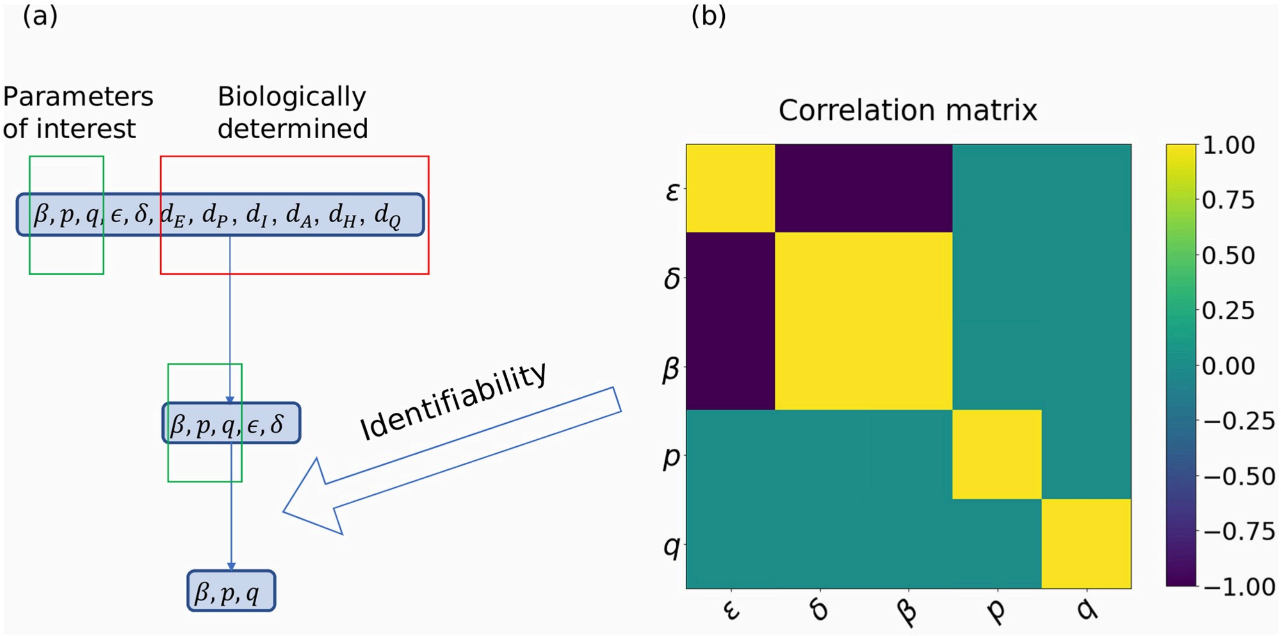

Fig 4. The procedure of choosing parameters to fit. (a) The procedure of determining parameters to fit. We fix d_E, d_P, d_I, d_A, d_H, d_Q because they are biologically determined, and then fix ϵ, δ due to the result from the correlation matrix analysis. (b) The correlation matrix of five parameters. Each colored off-diagonal cell represents the correlation between two parameters. Green means (almost) not statistically correlated while yellow/purple represents positively/negatively correlated, respectively.

https://doi.org/10.1371/journal.pcbi.1009334.g004

Table 5. Estimation of parameters, control reproduction number, and immunity threshold. The transmission rate β and the control reproduction number $\mathcal{R}_c$ change between different stages, indicating that local government policies in New York City and public holidays have a strong impact on the transmission dynamics of the pandemic.

	Meaning	Stage 1	Stage 2	Stage 3	Stage 4	Stage 5	Stage 6	Stage 7
β	Transmission rate	0.82	0.17	0.11	0.19	0.24	0.22	0.19
p	Hospitalization ratio	0.41	0.30	0.13	0.12	0.08	0.08	0.09
q	Death from hospital ratio	0.15	0.37	0.35	0.22	0.13	0.16	0.19
$\mathcal{R}_c$	Control reproduction number	4.55	0.97	0.60	1.07	1.35	1.24	1.06
IT	Immunity threshold	78.0%	0.0%	0.0%	6.4%	26.2%	19.3%	5.8%

https://doi.org/10.1371/journal.pcbi.1009334.t005

calculate the correlation matrix based on the parameters obtained in other stages; see S2 Fig. This indicates that ϵ, δ, and β are not practically identifiable. Two of them need to be fixed, while the rest and p, q can be fitted. In this paper, we fix ϵ, δ and fit β, p, q for the following reason: β, which represents the transmission rate, is highly affected by local government policy and does not have a stable and universal value compared to ϵ and δ. This means that the value of β could be very different across different datasets and is hard to determine a priori. Therefore, we fix ϵ and δ according to [36]. After that, the model becomes practically identifiable as shown in S3 Fig. A summary of the reasoning is shown in Fig 4(a). FIM is known to have limitations when their asymptotic assumptions are not satisfied, and they have limited ability to capture nonlinear dynamics. However, we observe that the fitted values of p and q in Table 4 are close to the results in Table 5 (when ϵ and δ are fixed), while the results for β are different. This indicates that FIM gives the correct result.

(IV) Sensitivity analysis

Variance-based sensitivity analysis, also called Sobol sensitivity analysis, is a global method that measures sensitivity across the whole input space. It decomposes the model's output variance into fractions that can be attributed to individual inputs or groups of inputs [18]. Suppose we are given a black box model:

$$y = f(\Theta), \tag{8}$$

where $y \in \mathbb{R}$ is the output and $\Theta = [\theta_1, \theta_2, \cdots, \theta_k] \in [0, 1]^k$ are independent and uniformly distributed uncertain inputs. If some components of Θ are not within [0, 1], we may transform Θ into the unit hypercube.

First-order sensitivity index measures the contribution to the output variance by a single input θ_i alone:

$$S_i(y) = \frac{\mathrm{Var}_{\theta_i}(\mathrm{E}_{\Theta_{\sim i}}(y|\theta_i))}{\mathrm{Var}(y)}, \tag{9}$$

where $\Theta_{\sim i} = [\theta_1, \cdots, \theta_{i-1}, \theta_{i+1}, \cdots, \theta_k]$. Total-order sensitivity index measures the contribution to the output variance by an input, including its first-order effect and all higher-order interactions with other inputs:

$$S_{Ti}(y) = 1 - \frac{\mathrm{Var}_{\Theta_{\sim i}}(\mathrm{E}_{x_i}(y|\Theta_{\sim i}))}{\mathrm{Var}(y)}. \tag{10}$$

Note that

$$S_{Ti}(y) \geq S_i(y) \tag{11}$$

by definition, and

$$\sum_{i=1}^{k} S_{Ti}(y) \geq 1 \tag{12}$$

since the interaction between θ_i and θ_j is counted in both $S_{Ti}(y)$ and $S_{Tj}(y)$.

Sensitivity analysis does not rely on any data. Instead, it analyzes the dependence relationship between the outputs and the inputs of a given model from the level of parametric equations when a specific initial condition to the system is given. Regarding the model in Fig 3, the cumulative infectious population I_{sum} in each stage of the pandemic is a function of the parameters β, p, and q (they are assumed constant during each period). So are the cumulative hospitalized population H_{sum} and the cumulative death population D_{sum}. Using Sobol's method, we obtain the first-order sensitivity and total-order sensitivity of each model output of interest (I_{sum}, H_{sum}, D_{sum}) with respect to each parameter (β, p, q). The ranges for β, p, and q are [0, 1]. For each model output, 8000 samples are generated using Saltelli's sampling scheme. The results are plotted in Fig 5. We can see that the cumulative cases I_{sum} does not depend on p or q. For the cumulative hospitalizations H_{sum}, β is the most important parameter while q does not have any impact. For the cumulative deaths D_{sum}, β is the most influential parameter as well. The qualitative relationship between the scale of sensitivity indices for different parameters is the same across all stages.

In the proposed model, the parameter β is the most important parameter for the projection of all I_{sum}, H_{sum}, and D_{sum}. Since p and q do not contribute to I_{sum}, our model may project I_{sum} even if p and q were inaccurate. Similarly, our model may project H_{sum} even if q were inaccurate.

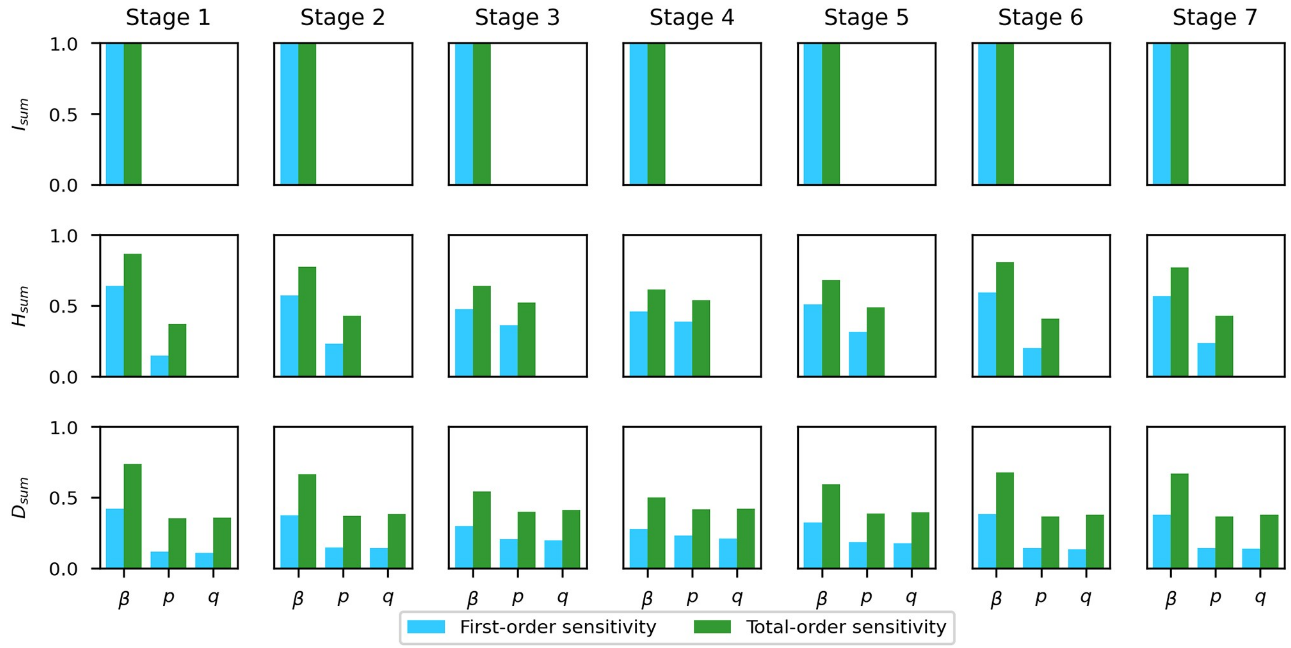

Fig 5. Sensitivity of each quantity of interest (I_{sum}, H_{sum},D_{sum}) with respect to each parameter (β, p, q). The parameter β is the most important parameter for all three quantities of interest in every stage of the pandemic. The parameter p has no influence on I_{sum}. The parameter q has no influence on I_{sum} or H_{sum}.

https://doi.org/10.1371/journal.pcbi.1009334.g005

(V) Model calibration

Estimation of the number of effective vaccinations

The vaccination data of COVID-19 in NYC are given in the form of the number of first and second doses administered from December 14, 2020 to February 4, 2021 (see Fig 2). We choose the Pfizer vaccine as a representative vaccine since it is the most commonly administered vaccine in the United States [56]. We use the parameters for the vaccine efficacy in the model from Pfizer's official data. The vaccine efficacy is 52% for only one dose and 95% for both doses [34]. As the vaccines are not 100% effective, in order to calculate the number of effectively vaccinated individuals that takes into account the vaccine efficacy, we use a weighted sum of the number of the first doses and second doses administered:

$$\begin{aligned}\text{Effectively vaccinated individuals per day} &= 0.52D_1 + (0.95 - 0.52)D_2 \\ &= 0.52(D_1 - D_2) + 0.95D_2,\end{aligned} \tag{13}$$

where D_1 is the number of individuals who receive the first dose and D_2 is the number of individuals who receive the second dose on that day. As the vaccines provide immunity 14 days after they are received, we remove the effectively vaccinated individuals from the susceptible (S) class and join them to the recovered (R) class 14 days after they have received the vaccines. To simplify the study, we approximate and project the number of daily effective vaccinations linearly with a cap of 20, 000 per day, which corresponds to a maximum capacity of about 40, 000 total doses per day; see Fig 6. This approximated and projected number is used as the time-dependent parameter v in our model (see Fig 3). Before any vaccine is effective, we have $v \equiv 0$.

Parameter estimation via simulated annealing

The data of daily cases, hospitalizations, and deaths of COVID-19 in NYC from February 29, 2020 to February 4, 2021 are given in Fig 2. We assume a constant population size of 8.399 million people in NYC and do not consider migration. Using the data with the model in Fig 3, initial values in Table 1, and parameters in Table 2, we fit the transmission rate β, hospitalization ratio p, and death from hospital ratio q within the range [0, 1]. The fitting is split into seven stages defined by the public policies described in Fig 2. In each stage, the parameters (β, p, q) are assumed to be constant as in Eq (4). We use simulated annealing, a global optimization algorithm, to search for the optimal parameter values in each stage. The objective is to minimize the following loss function:

$$\text{Loss} = \frac{\text{MSE}(I_{new}) + \text{MSE}(H_{new}) + \text{MSE}(D_{new})}{3} \tag{14}$$

in the region $\beta, p, q \in [0, 1]$, where "MSE" stands for mean squared error. The estimated final value in the previous stage is the initial value for the next stage. The results of all compartments are plotted in Figs 6 and 7. The time-dependent parameters (β, p, q) and the control reproduction number $\mathcal{R}_c$ are shown in Fig 8, Tables 5 and 6.

The control reproduction number $\mathcal{R}_c$, whose expression is in Eq (5), depends on six parameters: β, ϵ, δ, d_P, d_I, and d_A. Since ϵ, δ, d_P, d_I, and d_A are fixed as in Table 2, $\mathcal{R}_c$ is proportional to the transmission rate β; see Fig 8. We plot the evolution of $\mathcal{R}_c$ over time and overlay the scaled daily cases to demonstrate how the number of daily cases and $\mathcal{R}_c$ (or β) are related to each other. Before any closures took place on March 22, 2020, we had a high $\mathcal{R}_c$ with exponential growth of daily cases. Once the strict control measures were rolled out, $\mathcal{R}_c$ was considerably reduced below 1 along with a decline of daily cases. During the reopening Phases 1–4, $\mathcal{R}_c$

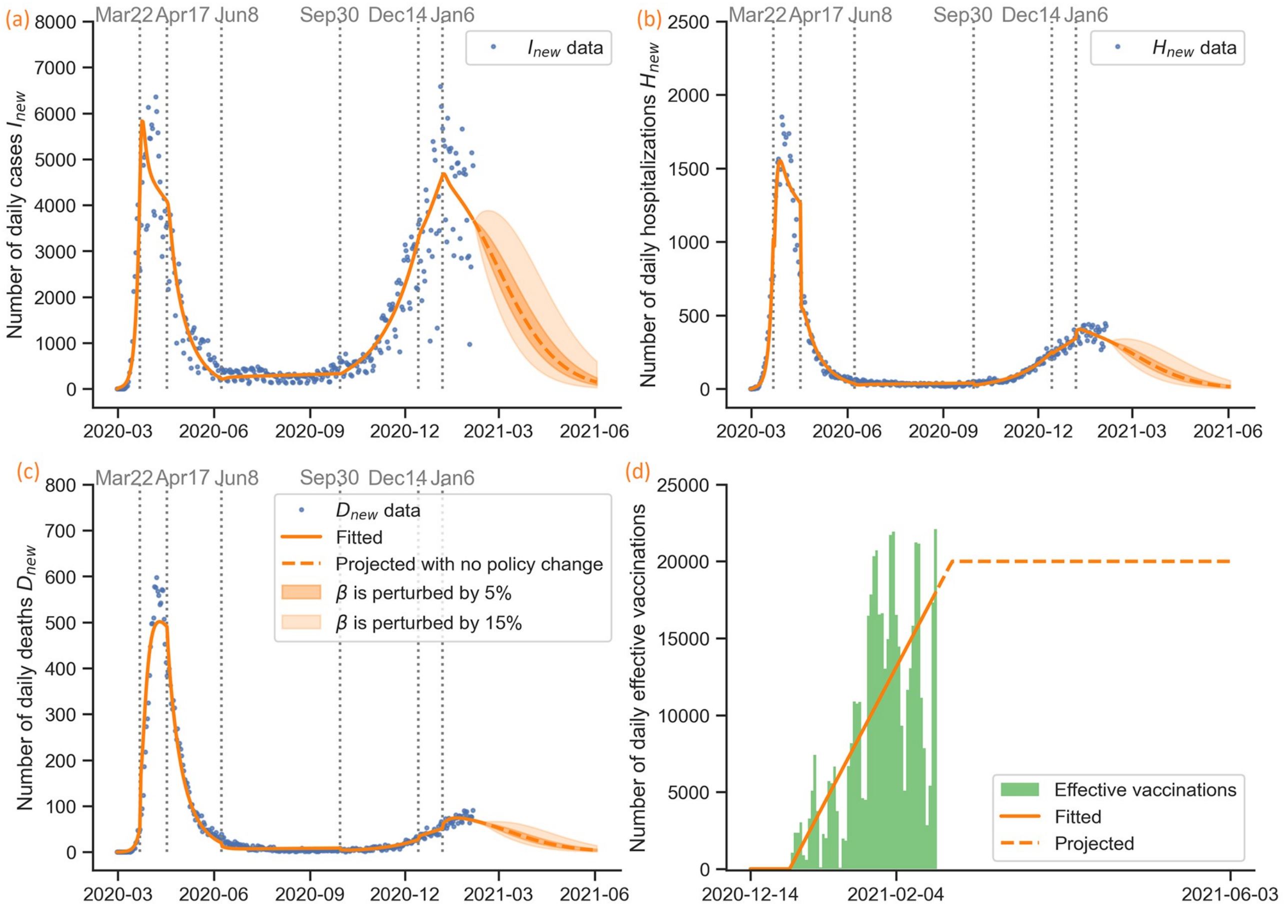

Fig 6. Estimation of daily cases, hospitalizations, deaths, and vaccinations in New York City. (a) Estimation of daily cases. (b) Estimation of daily hospitalizations. (c) Estimation of daily deaths. (d) We calculate the number of effective vaccinations as a weighted sum of the number of first and second doses administered as shown in Fig 2; we approximate the daily number of effective vaccinations linearly and assume it grows linearly until it reaches the maximum capacity of 20,000 per day.

https://doi.org/10.1371/journal.pcbi.1009334.g006

rose to around 1 with a stabilized number of daily cases. When indoor dining was reopened on September 30, 2020, $\mathcal{R}_c$ rose to above 1 with another wave of daily cases. After indoor dining was closed again on December 14, 2020, $\mathcal{R}_c$ decreased. After the end of holidays, $\mathcal{R}_c$ further decreased.

Bayesian posterior simulation via MCMC

We use the loss function Eq (14) as the negative log-likelihood of the posterior distribution of the parameters (β, p, q). We assume that each parameter's prior distribution is independent and uniformly distributed in [0, 1]. Using Markov chain Monte Carlo (MCMC) simulation, we may simulate the posterior distribution of each parameter associated with our approach. As before, the simulation is done within each stage where (β, p, q) are assumed to be constant. In each stage, four chains of 1000 samples are drawn with 200 burn-in samples in every chain. We initialize the chains at the estimation given by simulated annealing to speed up the algorithm; see S5–S11 Figs for the posterior distributions and the sampling processes. We can see that the chains are well-mixed, which implies the convergence of the sampling. The narrow posterior distributions indicate that our numerical algorithm is robust, the quantity of data is

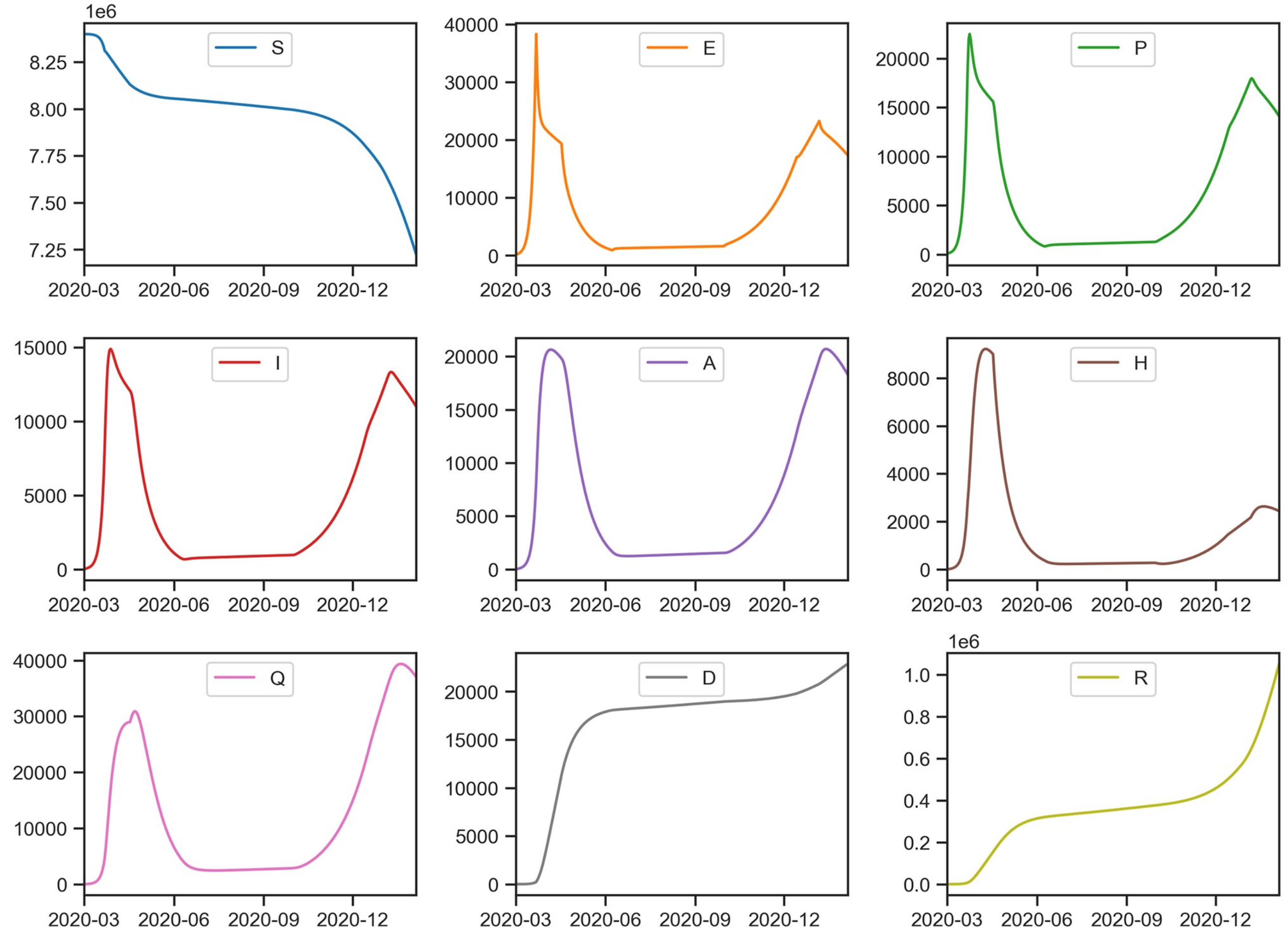

Fig 7. Estimation of the unobserved dynamics in all the model compartments (*S*,*E*, *P*, *I*, *A*, *H*, *Q*, *D*, *R*). The number of susceptible individuals (*S*) drops significantly as the number of cases hikes after December 2020.

https://doi.org/10.1371/journal.pcbi.1009334.g007

sufficient for our approach of minimizing the loss function (14), and the parameters are indeed close to constant in every stage. As a result, the parameter estimation is reliable.

(VI) Model robustness analysis

Using the fitted parameter values in Table 5, we perform the Monte Carlo simulation to check the robustness of our model to perturbations, adapting ideas from [17, 57]. The model robustness analysis is another form of practical identifiablity analysis, which can be seen as a complement to the FIM-based approaches. The computational cost of this method is high due to its Monte Carlo nature. However, it is necessary because FIM-based methods are known to have limitations, as discussed in (III).

We first multiply the daily increase in the calibrated data (a subset of $\{I_{sum}, H_{sum}, D_{sum}\}$) by independent and identically distributed Gaussian random noise of mean 1 and standard deviation σ to generate a new dataset, which looks like our original dataset with measurement error. Then, we estimate the parameters by fitting the model to the artificially generated dataset and compare the result with the parameter values obtained in Table 5. The same procedure is repeated for $M = 1000$ times, and we compute the average error between the parameter values estimated from the original and the generated datasets. The quantity we obtained is named

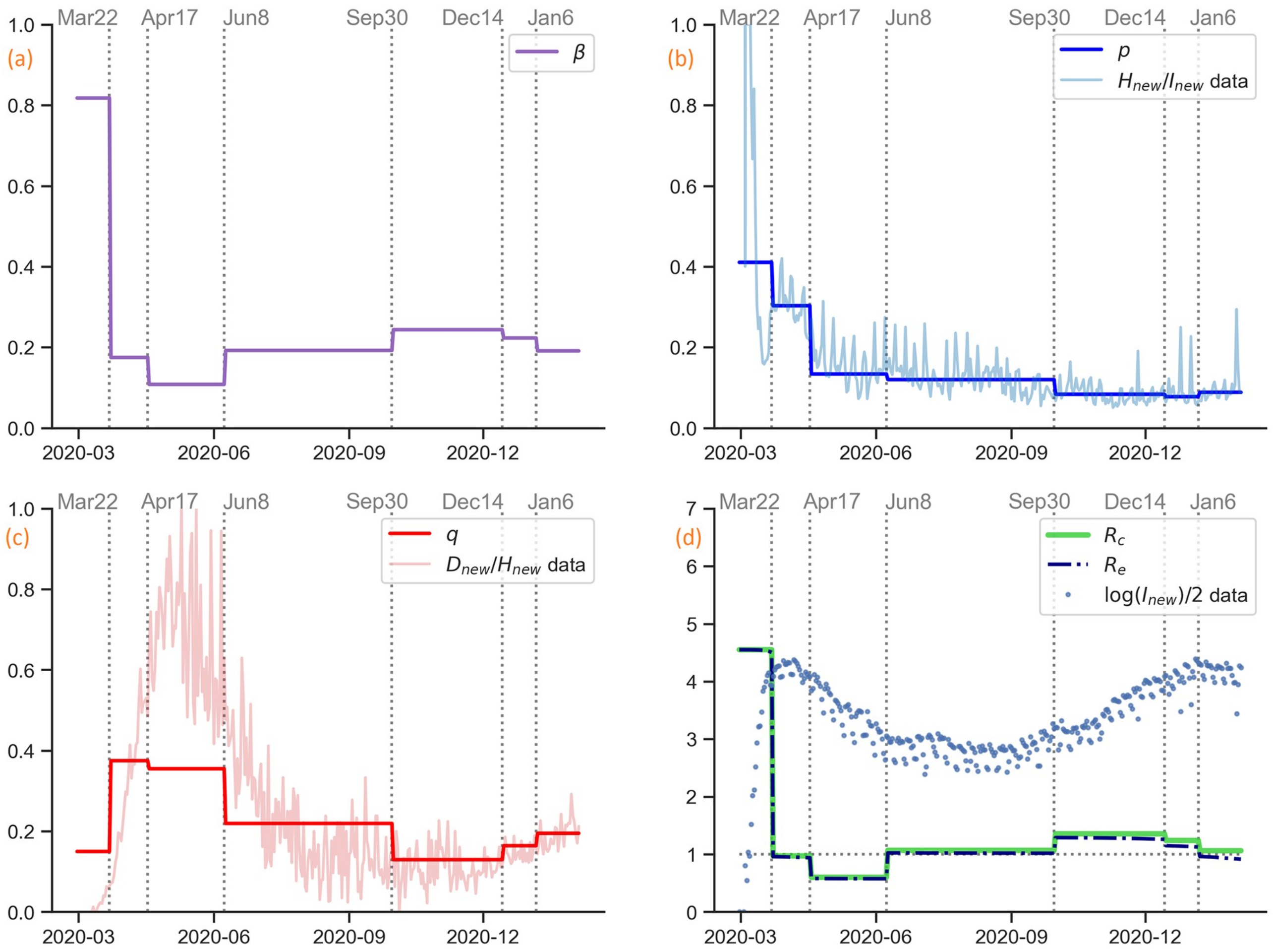

Fig 8. Estimation of parameters and reproduction numbers. (a) Estimated time-dependent transmission rate $\beta(t)$. (b) Estimated time-dependent hospitalization ratio $p(t)$, compared with daily hospitalizations over daily cases calculated from the raw data. (c) Estimated time-dependent death from hospital ratio $q(t)$, compared with daily deaths over daily hospitalizations calculated from the raw data. (d) Estimated control reproduction number $\mathcal{R}_c$ and effective reproduction number $\mathcal{R}_e$ calculated by the estimated parameters, compared with 1/2 of the logarithm of daily cases.

https://doi.org/10.1371/journal.pcbi.1009334.g008

average relative error (ARE):

$$\text{ARE}(\theta_i) = \frac{1}{M\sigma}\sum_{j=1}^{M}\left|\frac{\hat{\theta}_i^{(j)} - \theta_i}{\theta_i}\right|, \tag{15}$$

where θ_i is the fitted value of the ith parameter (i.e., β, p, q) on the original dataset, and $\hat{\theta}_i^{(j)}$ is the fitted value of the ith parameter on the jth generated dataset.

Table 6. Percentage changes of parameters and control reproduction number between contiguous stages. The stay-at-home order in Stage 2, mask mandate in Stage 3, closing of indoor dining and starting of vaccination in Stage 6, and end of the holidays in Stage 7 lead to decreases in the transmission rate β and the reproduction number $\mathcal{R}_c$. The four-phase reopening in Stage 4 and reopening of indoor dining in Stage 5 lead to increases in β and $\mathcal{R}_c$.

	Meaning	Stage 1–2	Stage 2–3	Stage 3–4	Stage 4–5	Stage 5–6	Stage 6–7
β	Transmission rate	−78.7%	−38.3%	+ 78.0%	+ 26.8%	−8.5%	−14.4%
p	Hospitalization ratio	−26.2%	−55.9%	−10.4%	−30.4%	−7.0%	+ 13.9%
q	Death from hospital ratio	+ 150.4%	−5.3%	−38.2%	−41.0%	+ 26.7%	+ 18.8%
$\mathcal{R}_c$	Control reproduction number	−78.7%	−38.3%	+ 78.0%	+ 26.8%	−8.5%	−14.4%

https://doi.org/10.1371/journal.pcbi.1009334.t006

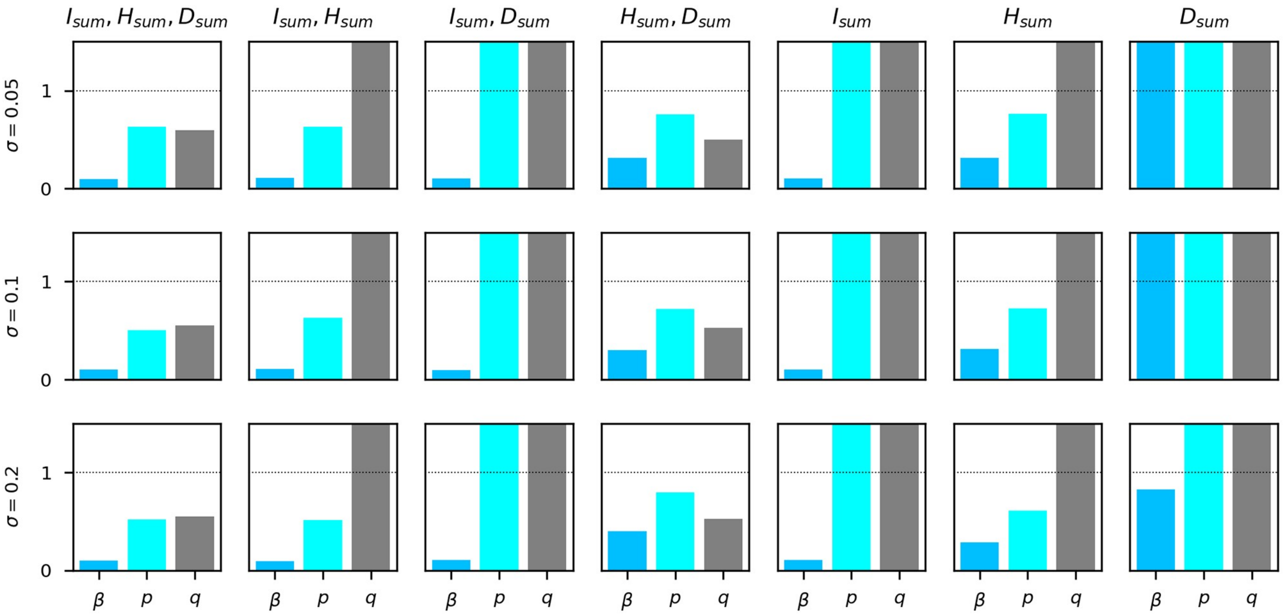

Fig 9. Average Relative Error (ARE) of (β, p, q) in different observable settings. Each row corresponds to a standard deviation level of random noise multiplied to the observables. Each column represents an observable setting. When (I_{sum}, H_{sum}, D_{sum}) or (H_{sum}, D_{sum}) are given, ARE is lower than the threshold 1. Therefore, our model is robust to noise in the NYC dataset. In the rest of of the missing observable cases, our model would not be robust to perturbations, which is consistent with the structural identifiability result.

https://doi.org/10.1371/journal.pcbi.1009334.g009

When the parameters are piecewise constant and fitted separately, we define the overall ARE of that parameter to be the largest ARE calculated in every stage. For example, in this paper, ARE(β) = $\max_{s\in\{1,\cdots,7\}}$ ARE(β_s). Finally, we define the maximum average relative error (MARE) of the model to be the largest ARE of all the model parameters:

$$\mathrm{MARE} = \max_{i\in\{1,\cdots,k\}} \mathrm{ARE}(\theta_i), \tag{16}$$

where k is the total number of parameters to be estimated. If MARE < 1, we say the model is robust to perturbation. The algorithm is detailed in S5 Text.

The first column in Fig 9 shows that when (I_{sum}, H_{sum}, D_{sum}) or (H_{sum}, D_{sum}) are given, our model is robust to noise, which justifies the fitting result (since β, p, q are also identifiable) and provides a theoretical backup for the projection in the next section. The other columns in Fig 9 show that even when H_{sum} or D_{sum} is missing, the model would not be robust to perturbation, which is consistent with the structural identifiability result.

(VII) Projection with uncertainties and scenarios

The situation in NYC evolves day by day. The city reinstated indoor dining restrictions in mid-December due to the steady increase in the virus incidence. The ever-changing policies add a high level of uncertainty to any long term projection we can make. Here, we explore our model's ability to project the number of daily cases, hospitalizations, and deaths in the city with uncertainty.

The MCMC simulation provides us with a way to quantify uncertainty. We may sample from the posterior distribution of the parameters in the last stage (see S11 Fig) and run the model after that to obtain a distribution of the projected daily cases, hospitalizations, and deaths. However, this approach assumes that the situation remains the same after the last stage, which may not be the case. There might be policy changes or other events. As a result,

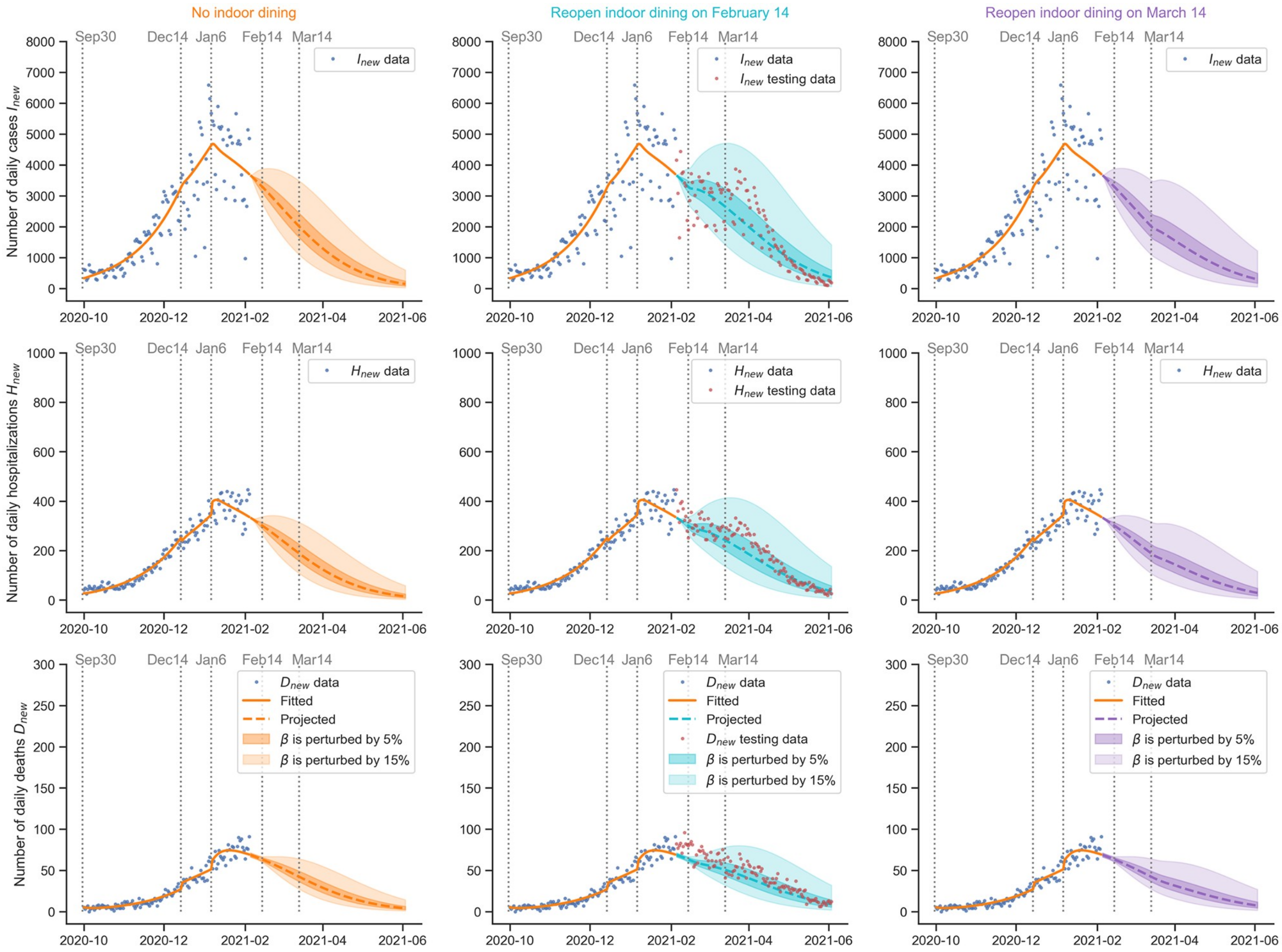

Fig 10. Projection of daily cases, hospitalizations, and deaths in New York City with uncertainties and scenarios. Reopening scenarios on February 14 and March 14 are considered. An increase in infectious, hospitalized, and deceased population is expected if the restaurants are reopened in the same way as Stage 5 (September 30, 2020 to December 14, 2020). Postponing the reopening of restaurants from February 14 to March 14 may reduce the number of infectious, hospitalized, and deceased individuals. The actual situation might vary depending on the details and implementations of the actual indoor dining policies that take place in 2021. **Remarks**: The projections were made and the paper was submitted in February. When updating the paper in June, we overlaid the new data of daily cases, hospitalizations, and deaths from February to June as the testing data. Indoor dining was actually reopened on February 14.

https://doi.org/10.1371/journal.pcbi.1009334.g010

we perturb the transmission rate β by a percentage to reflect future policy changes or other events. We sample from the posterior distribution of the parameters (β, p, q) in the last stage. Then, we multiply β by a random number drawn from a uniform distribution ($\mathcal{U}(0.95, 1.05)$ or $\mathcal{U}(0.85, 1.15)$). In other words, we perturb β by 5% or 15%. As a reference, see Table 6 for the historical changes of β between contiguous stages. We run the model with the initial value as the ending value of the last stage to project daily cases, hospitalizations, and deaths. After repeating 4000 times, we obtain a distribution of the daily cases, hospitalizations, and deaths at every timestamp after the last stage. Then, the 95% confidence interval at every timestamp is plotted.

In Fig 10, we consider three different scenarios: no indoor dining, reopening indoor dining on February 14, and reopening indoor dining on March 14. The uncertainties given by the perturbed β's are plotted in each scenario. For indoor dining, we assume it is a 25% reopening, which is the same as what happened in Stage 5 (September 30, 2020 to December 14, 2020).

We multiply β by β_5/β_6 to represent the change of the transmission rate caused by the reopening of indoor dining. An increase in infectious, hospitalized, and deceased population is expected if the restaurants are reopened. Postponing the reopening of restaurants from February 14 to March 14 may reduce the number of infectious, hospitalized, and deceased individuals. The actual situation might vary depending on the details and implementations of the actual indoor dining policies that take place in 2021.

Discussion

The COVID-19 epidemic is an unprecedented worldwide public health challenge, especially in densely populated areas such as New York City (NYC). Epidemiological models can provide the dynamic evolution of a pandemic but they are based on many assumptions and parameters that have to be adjusted over the time when the pandemic lasts. However, the available data might not be sufficient to identify the model parameters and hence infer the unobserved dynamics. This is typical of any past epidemics or pandemics, and hence a systematic integrated framework is required to make existing or modified models useful for designing health policies.

To this end and after studying the current pandemic for almost a year, we have designed a general framework as shown in Fig 1 for building a trustworthy data-driven epidemiological model, which constructs a workflow to integrate data acquisition and event timeline, model development, identifiability analysis, sensitivity analysis, model calibration, model robustness analysis, and projection with uncertainties in different scenarios. The proposed general framework can provide guidance on how to build the appropriate epidemiological model based on the available data in a specific region. The proposed framework can help to assess the structural and practical identifiability of model parameters so that model parameters can be uniquely estimated, and the calibrated model can make more robust and reliable projections that can be used for evaluating the effect of vaccination and various other scenarios.

In particular, we apply this framework to first endow the SEIR model with more compartments, and subsequently we extend to include vaccination, with the objective of projecting the transmission dynamics of COVID-19 in NYC under vaccination and different safety measures relaxation scenarios. Based on the proposed general framework, we first acquire data from the NYC's government's website and look for major intervention events that could affect the transmission dynamics of the pandemic. We then develop a mathematical model that describes the COVID-19 infection's biological characteristics by extending the SEIR model to include presymptomatic, asymptomatic, isolated, hospitalized, and deceased individuals. This model takes advantage of all the epidemiological data available from the COVID-19 outbreak in NYC by fitting hospitalizations and disease-related deaths in addition to the daily cases. Furthermore, we incorporate the effects of intervention strategies in the outbreak's evolution by including time-dependent parameters to capture these variations.

Given a model and epidemiological data, this framework addresses the problem of identifying which parameters can be inferred accurately. We perform two types of identifiability analysis, structural identifiability and practical identifiability analysis, to address this problem. From the structural identifiability analysis, we conclude that five parameters (β, p, q, ϵ, δ) of the proposed model are structurally globally identifiable when daily cases (I_{new}), hospitalizations (H_{new}) and deaths (D_{new}) are provided, which is the case for the NYC dataset. However, when H_{sum} or D_{sum} is not available, one cannot get a trustworthy estimation of these parameters since at least one of the model parameters would be non-identifiable. For the purpose of reliable parameter estimation, one should utilize all the provided data in NYC.

The Fisher correlation matrix method enables us to determine that two out of five parameters need to be fixed due to practical non-identifiability, even if all the data (I_{new}, H_{new}, D_{new})

in the NYC dataset are given. Therefore, we use the values of δ and ϵ provided by the CDC pandemic planning scenarios, and once we fix these two parameters (ϵ, δ), the other three parameters are practically identifiable. For some other cities, however, it can be challenging to maintain careful records of infected, hospitalized, and deceased individuals in an ongoing epidemic. The robustness analysis allows us to conclude that we can still project some variables with a degree of accuracy despite missing infectious data since in this case the model is still robust to noise.

As a result, we fit three parameters (β, p, q) given three observables (I_{new}, H_{new}, D_{new}). Sensitivity analysis demonstrates that the parameter β is the most important parameter for the projection of all I_{sum}, H_{sum}, and D_{sum}. Since p and q do not contribute to I_{sum}, our model may project I_{sum} even if p and q were inaccurate. Similarly, our model may project H_{sum} even if q were inaccurate.

We observe that the proposed data-driven epidemiological model can uniquely estimate the model parameters. The fitted daily cases, hospitalizations, and deaths match with the data from the NYC's government's website. In addition, we employ Monte Carlo simulations to quantify the uncertainties in the parameters and project under uncertainties. We employ the calibrated data-driven model to study the effects of the timing of reopening indoor dining. The projection results indicate that postponing the reopening of restaurants from February 14 to March 14 may reduce the number of infectious, hospitalized, and deceased individuals. Such a projection can be readily updated as new data are accumulated. The actual situation might vary depending on the details and implementations of the actual indoor dining policies that take place in 2021 and corresponding updates are required.

Assumptions and limitations of the model

This study has some limitations resulting from the model's structure and the consideration of identifiability. Therefore, the results are subject to several simplifying assumptions.

- As a modified SEIR model, our model assumes that there is no migration into or out of NYC, no births or other deaths besides COVID-related deaths, and the population is well mixed.
- We are fitting to the observed COVID-19 cases without correcting for undetected symptomatic infections. Although we do not include the undetected infections in the main model in the manuscript, we explore an alternative model taking into account the ascertainment ratio in NYC in S6 Text.
- All symptomatic infections will either self-isolate or become hospitalized. In the case of NYC, due to the NYC Test & Trace Corps, a high percentage of symptomatic and detected individuals adhered to self quarantine [58]. The Global Health Governance Program reports that in-person tracers locate approximately 80% of people at home [58, 59].
- Hospitalized individuals do not transmit the disease. The percentage of hospital-acquired infections is highly dependent on the location and varies widely. Hospital-acquired cases reached 16.2% in England [60]; however, in a study in US hospitals, such as the one by Rhee et al. [61], the incidence of hospital-acquired COVID-19 is low and negligible. Rhee et al. studied all patients admitted to Brigham and Women's Hospital (Boston, Massachusetts) between March 7, 2020 and May 30, 2020. They found 1 COVID-19 patient deemed to be hospital-acquired and another one deemed likely to be hospital-acquired, though with no known exposures [61].
- All deceased individuals have gone through the hospital (H) compartment. The NYC's government's official data classify confirmed deaths as deaths within 60 days of a positive

molecular test, i.e., they have gone through the symptomatic (I) compartment, but they might not have been admitted to hospital in practice. The daily deaths reported may be from non-hospital settings such as nursing homes. Our model assumes that all deaths are from hospital for the sake of identifiability and is based on the situation in NYC, where the non-hospital deaths are of very low proportion within all the reported deaths. As a result, the death from hospital ratio (q) might be slightly over estimated.

- Initial conditions are calculated directly from the available data.
- Date ranges that determine the different stages are fixed and based on state/city policy changes.
- If we were to implement this model at the early stage of a pandemic, we might expect lower accuracy due to uncertainties about the values of the fixed parameters.

Supporting information

S1 Text. Model development.
(PDF)

S2 Text. Parameter settings.
(PDF)

S3 Text. Basic and control reproduction number of the model in Fig 3.
(PDF)

S4 Text. Definition and algorithm for identifiability analysis.
(PDF)

S5 Text. Definition and algorithm for model robustness.
(PDF)

S6 Text. Alternative setups of the model.
(PDF)

S1 Table. Data to estimate d_H.
(PDF)

S2 Table. Structural identifiability of the model with H as an observable. Different from Table 3, we assume H instead of H_{sum} is observed as data.
(PDF)

S1 Fig. Raw model.
(PDF)

S2 Fig. Correlation matrix of β, p, q, δ, and ϵ in different stages.
(PDF)

S3 Fig. Correlation matrix of β, p, and q in the setting of different observables.
(PDF)

S4 Fig. Fitting and projection with the data given as (I_{new}, H_{new}), (H_{new}, D_{new}), or (I_{new}).
(PDF)

S5 Fig. MCMC simulation in Stage 1.
(PDF)

S6 Fig. MCMC simulation in Stage 2.
(PDF)

S7 Fig. MCMC simulation in Stage 3.
(PDF)

S8 Fig. MCMC simulation in Stage 4.
(PDF)

S9 Fig. MCMC simulation in Stage 5.
(PDF)

S10 Fig. MCMC simulation in Stage 6.
(PDF)

S11 Fig. MCMC simulation in Stage 7.
(PDF)

S12 Fig. Fitting and projection of (I_{new}, H_{new}, D_{new}) if we shift the policy stay-at-home order at the beginning of the outbreak from March 22 to March 17.
(PDF)

S13 Fig. Plot of the time-dependent ascertainment ratio δ.
(PDF)

S14 Fig. Fitting and projection of (I_{new}, H_{new}, D_{new}) with time-dependent ascertainment ratio δ.
(PDF)

S15 Fig. Fitting and projection of (I_{new}, H_{new}, D_{new}) with time-dependent ascertainment ratio δ and alternate initial value.
(PDF)

Author Contributions

Conceptualization: Sheng Zhang, Joan Ponce, Guang Lin, George Karniadakis.

Data curation: Sheng Zhang.

Formal analysis: Sheng Zhang, Zhen Zhang.

Funding acquisition: Guang Lin, George Karniadakis.

Investigation: Sheng Zhang, Joan Ponce, Zhen Zhang.

Methodology: Sheng Zhang, Joan Ponce, Zhen Zhang.

Project administration: Sheng Zhang, Joan Ponce, Guang Lin, George Karniadakis.

Resources: Sheng Zhang, Joan Ponce, Zhen Zhang, Guang Lin, George Karniadakis.

Software: Sheng Zhang, Zhen Zhang.

Supervision: Guang Lin, George Karniadakis.

Validation: Sheng Zhang, Zhen Zhang.

Visualization: Sheng Zhang, Zhen Zhang.

Writing – original draft: Sheng Zhang, Joan Ponce, Zhen Zhang.

Writing – review & editing: Sheng Zhang, Joan Ponce, Zhen Zhang, Guang Lin, George Karniadakis.

References

1. Zheng W. Total Variation Regularization for Compartmental Epidemic Models with Time-varying Dynamics. arXiv preprint arXiv:200400412. 2020; p. 1–11.
2. Roda WC, Varughese MB, Han D, Li MY. Why is it difficult to accurately predict the COVID-19 epidemic? Infectious Disease Modelling. 2020; 5:271–281. https://doi.org/10.1016/j.idm.2020.03.001 PMID: 32289100
3. Lourenco J, Paton R, Ghafari M, Kraemer M, Thompson C, Simmonds P, et al. Fundamental principles of epidemic spread highlight the immediate need for large-scale serological surveys to assess the stage of the SARS-CoV-2 epidemic. MedRxiv. 2020; p. 1–7.
4. Maier BF, Brockmann D. Effective containment explains subexponential growth in recent confirmed COVID-19 cases in China. Science. 2020; 368(6492):742–746. https://doi.org/10.1126/science.abb4557 PMID: 32269067
5. Giordano G, Blanchini F, Bruno R, Colaneri P, Di Filippo A, Di Matteo A, et al. Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy. Nature Medicine. 2020; p. 1–6. https://doi.org/10.1038/s41591-020-0883-7 PMID: 32322102
6. Gaeta G. A simple SIR model with a large set of asymptomatic infectives. arXiv preprint arXiv:200308720. 2020; p. 1–23.
7. Shi P, Cao S, Feng P. SEIR transmission dynamics model of 2019 nCoV coronavirus with considering the weak infectious ability and changes in latency duration. MedRxiv. 2020; p. 1–5.
8. Zha WT, Pang FR, Zhou N, Wu B, Liu Y, Du YB, et al. Research about the optimal strategies for prevention and control of varicella outbreak in a school in a central city of China: Based on an SEIR dynamic model. Epidemiology & Infection. 2020; 148.
9. He S, Peng Y, Sun K. SEIR modeling of the COVID-19 and its dynamics. Nonlinear Dynamics. 2020; 101(3):1667–1680. https://doi.org/10.1007/s11071-020-05743-y PMID: 32836803
10. López L, Rodo X. A modified SEIR model to predict the COVID-19 outbreak in Spain and Italy: simulating control scenarios and multi-scale epidemics. Results in Physics. 2021; 21:103746. https://doi.org/10.1016/j.rinp.2020.103746 PMID: 33391984
11. Yang Z, Zeng Z, Wang K, Wong SS, Liang W, Zanin M, et al. Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. Journal of Thoracic Disease. 2020; 12(3):165. https://doi.org/10.21037/jtd.2020.02.64 PMID: 32274081
12. Rodriguez-Fernandez M, Mendes P, Banga JR. A hybrid approach for efficient and robust parameter estimation in biochemical pathways. Biosystems. 2006; 83(2-3):248–265. https://doi.org/10.1016/j.biosystems.2005.06.016 PMID: 16236429
13. Rodriguez-Fernandez M, Egea JA, Banga JR. Novel metaheuristic for parameter estimation in nonlinear dynamic biological systems. BMC bioinformatics. 2006; 7(1):483. https://doi.org/10.1186/1471-2105-7-483 PMID: 17081289
14. Massonis G, Banga JR, Villaverde AF. Structural identifiability and observability of compartmental models of the COVID-19 pandemic. Annual reviews in control.2020; p. 1–19. https://doi.org/10.1016/j.arcontrol.2020.12.001 PMID: 33362427
15. Gallo L, Frasca M, Latora V, Russo G. Lack of practical identifiability may hamper reliable predictions in COVID-19 epidemic models. arXiv preprint arXiv:201200443. 2020; p. 1–33.
16. Lee C, Li Y, Kim J. The susceptible-unidentified infected-confirmed (SUC) epidemic model for estimating unidentified infected population for COVID-19. Chaos, Solitons & Fractals. 2020; 139:110090. https://doi.org/10.1016/j.chaos.2020.110090 PMID: 32834625
17. Tuncer N, Le TT. Structural and practical identifiability analysis of outbreak models. Mathematical Biosciences. 2018; 299:1–18. https://doi.org/10.1016/j.mbs.2018.02.004 PMID: 29477671
18. Sobol IM. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. Mathematics and computers in simulation. 2001; 55(1-3):271–280. https://doi.org/10.1016/S0378-4754(00)00270-6
19. WHO. Coronavirus disease 2019 (COVID-19) Situation Report—51; 2020. https://apps.who.int/iris/bitstream/handle/10665/331475/nCoVsitrep11Mar2020-eng.pdf?sequence=1&isAllowed=y.
20. Yan H, Sgueglia K. New York's first case of coronavirus is a health care worker, and officials say more cases are 'inevitable'; 2020. Available from: https://edition.cnn.com/2020/03/02/us/new-york-coronavirus-first-case/index.htm.

21. West MG. First Case of Coronavirus Confirmed in New York State; 2020. Available from: https://www.wsj.com/articles/first-case-of-coronavirus-confirmed-in-new-york-state-11583111692.

22. Governor A M Cuomo. Governor Cuomo Signs the'New York State on PAUSE' Executive Order; 2020. https://www.governor.ny.gov/news/governor-cuomo-signs-new-york-state-pause-executive-order.

23. NYC Health Department. NYC Coronavirus Disease 2019 (COVID-19) Data; 2020. https://github.com/nychealth/coronavirus-data#nyc-coronavirus-disease-2019-covid-19-data.

24. Holmdahl I, Buckee C. Wrong but useful–what covid-19 epidemiologic models can and cannot tell us. New England Journal of Medicine. 2020; 383(4):303–305. https://doi.org/10.1056/NEJMp2016822 PMID: 32412711

25. Governor A M Cuomo. At Novel Coronavirus Briefing, Governor Cuomo Declares State of Emergency to Contain Spread of Virus; 2020. https://www.governor.ny.gov/news/novel-coronavirus-briefing-governor-cuomo-declares-state-emergency-contain-spread-virus.

26. Governor A M Cuomo. Amid Ongoing COVID-19 Pandemic, Governor Cuomo Issues Executive Order Requiring All People in New York to Wear Masks or Face Coverings in Public; 2020. https://www.governor.ny.gov/news/amid-ongoing-covid-19-pandemic-governor-cuomo-issues-executive-order-requiring-all-people-new.

27. CDC. Considerations for Wearing Masks: Help Slow the Spread of COVID-19;2020. https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/cloth-face-cover-guidance.html(03-April-2020).

28. Governor A M Cuomo. NY Forward: A Guide to Reopening New York & Building Back Better; 2020. https://www.governor.ny.gov/sites/governor.ny.gov/files/atoms/files/NYForwardReopeningGuide.pdf.

29. New York State Government. Phase One Industry; 2020. https://forward.ny.gov/phase-one-industries.

30. New York State Government. Phase Two Industry; 2020. https://forward.ny.gov/phase-two-industries.

31. Andrew S, Myles M. NYC Indoor Dining to Resume Sept. 30 With Heavy Restrictions, Cuomo Says; 2020. https://www.nbcnewyork.com/news/local/nyc-casinos-malls-reopen-today-cuomo-and-de-blasio-face-lawsuit-over-indoor-dining/2607750.

32. Gold M. Indoor Dining Will Shut Down in New York City Again; 2020. https://www.nytimes.com/2020/12/11/nyregion/indoor-dining-nyc.html.

33. Luis FS, Joseph G. 1st Vaccination in U.S. Is Given in New York, Hard Hit in Outbreak's First Days; 2020. https://www.nytimes.com/2020/12/14/nyregion/coronavirus-vaccine-new-york.html.

34. Polack FP, Thomas SJ, Kitchin N, Absalon J, Gurtman A, Lockhart S, et al. Safety and efficacy of the BNT162b2 mRNA Covid-19 vaccine. New England Journal of Medicine. 2020; 383(27):2603–2615. https://doi.org/10.1056/NEJMoa2034577 PMID: 33301246

35. Baden LR, El Sahly HM, Essink B, Kotloff K, Frey S, Novak R, et al. Efficacy and safety of the mRNA-1273 SARS-CoV-2 vaccine. New England Journal of Medicine. 2020; p. 403–416. https://doi.org/10.1056/NEJMoa2035389 PMID: 33378609

36. CDC. COVID-19 Pandemic Planning Scenarios; 2020. https://www.cdc.gov/coronavirus/2019-ncov/hcp/planning-scenarios.html(10-July-2020).

37. Hao X, Cheng S, Wu D, Wu T, Lin X, Wang C. Reconstruction of the full transmission dynamics of COVID-19 in Wuhan. Nature. 2020; 584(7821):420–424. https://doi.org/10.1038/s41586-020-2554-8 PMID: 32674112

38. Cheng HY, Jian SW, Liu DP, Ng TC, Huang WT, Lin HH, et al. Contact tracing assessment of COVID-19 transmission dynamics in Taiwan and risk at different exposure periods before and after symptom onset. JAMA internal medicine. 2020; 180(9):1156–1163. https://doi.org/10.1001/jamainternmed.2020.2020 PMID: 32356867

39. Byambasuren O, Cardona M, Bell K, Clark J, McLaws ML, Glasziou P. Estimating the extent of true asymptomatic COVID-19 and its potential for community transmission: systematic review and meta-analysis. Available at SSRN 3586675. 2020; p. 1–14.

40. Mizumoto K, Kagaya K, Zarebski A, Chowell G. Estimating the asymptomatic proportion of coronavirus disease 2019 (COVID-19) cases on board the Diamond Princess cruise ship, Yokohama, Japan, 2020. Eurosurveillance. 2020; 25(10):2000180. https://doi.org/10.2807/1560-7917.ES.2020.25.10.2000180

41. Bai Y, Yao L, Wei T, Tian F, Jin DY, Chen L, et al. Presumed asymptomatic carrier transmission of COVID-19. JAMA. 2020; 323(14):1406–1407. https://doi.org/10.1001/jama.2020.2565 PMID: 32083643

42. Nishiura H, Kobayashi T, Miyama T, Suzuki A, Jung Sm, Hayashi K, et al. Estimation of the asymptomatic ratio of novel coronavirus infections (COVID-19). International journal of infectious diseases. 2020; 94:154. https://doi.org/10.1016/j.ijid.2020.03.020 PMID: 32179137

43. Wu JT, Leung K, Leung GM. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. The Lancet. 2020; 395(10225):689–697. https://doi.org/10.1016/S0140-6736(20)30260-9 PMID: 32014114

44. Furukawa NW, Brooks JT, Sobel J. Evidence supporting transmission of severe acute respiratory syndrome coronavirus 2 while presymptomatic or asymptomatic. Emerging infectious diseases. 2020; 26 (7). https://doi.org/10.3201/eid2607.201595

45. Kimball A, Hatfield KM, Arons M, James A, Taylor J, Spicer K, et al. Asymptomatic and presymptomatic SARS-CoV-2 infections in residents of a long-term care skilled nursing facility—King County, Washington, March 2020. Morbidity and Mortality Weekly Report. 2020; 69(13):377. https://doi.org/10.15585/mmwr.mm6913e1 PMID: 32240128

46. Anderson RM, May RM. The population dynamics of microparasites and their invertebrate hosts. Philosophical Transactions of the Royal Society of London B, Biological Sciences. 1981; 291(1054):451–524. https://doi.org/10.1098/rstb.1981.0005

47. Horwitz L, Jones SA, Cerfolio RJ, Francois F, Greco J, Rudy B, et al. Trends in Covid-19 risk-adjusted mortality rates in a single health system. medRxiv. 2020; p. 1–5.

48. Petrilli CM, Jones SA, Yang J, Rajagopalan H, O'Donnell LF, Chernyak Y, et al. Factors associated with hospitalization and critical illness among 4103 patients with COVID-19 disease in New York City. MedRxiv. 2020; p. 1–25.

49. Bellman R, Åström KJ. On structural identifiability. Mathematical Biosciences. 9131970; 7(3-4):329–339. https://doi.org/10.1016/0025-5564(70)90132-X

50. Rothenberg TJ. Identification in parametric models. Econometrica: Journal of the Econometric Society. 1971; p. 577–591. https://doi.org/10.2307/1913267

51. Glover K, Willems J. Parametrizations of linear dynamical systems: Canonical forms and identifiability. IEEE Transactions on Automatic Control.9181974; 19(6):640–646. https://doi.org/10.1109/TAC.1974.1100711

52. Thowsen A. Identifiability of dynamic systems. International Journal of Systems Science. 1978; 9 (7):813–825. https://doi.org/10.1080/00207727808941738

53. Reid J. Structural identifiability in linear time-invariant systems. IEEE Transactions on Automatic Control. 1977; 22(2):242–246. https://doi.org/10.1109/TAC.1977.1101474

54. Hong H, Ovchinnikov A, Pogudin G, Yap C. SIAN: software for structural identifiability analysis of ODE models. Bioinformatics. 2019; 35(16):2873–2874. https://doi.org/10.1093/bioinformatics/bty1069 PMID: 30601937

55. Ligon TS, Fröhlich F, Chişs OT, Banga JR, Balsa-Canto E, Hasenauer J. GenSSI2.0: multi-experiment structural identifiability analysis of SBML models. Bioinformatics. 2018; 34(8):1421–1423. https://doi.org/10.1093/bioinformatics/btx735 PMID: 29206901

56. Our World in Data Organization Our world in data. 2021. https://ourworldindata.org/grapher/covid-vaccine-doses-by-930manufacturer?tab=table.

57. Miao H, Xia X, Perelson AS, Wu H. On identifiability of nonlinear ODE models and applications in viral dynamics. SIAM review. 2011; 53(1):3–39. https://doi.org/10.1137/090757009 PMID: 21785515

58. Hospitals NH. NYC Test & Trace Corps; 2021. https://www.nychealthandhospitals.org/test-and-trace/take-care/.

59. Patel J, Fernandes G, Anchuri K. SELF-ISOLATION: SUPPORT, MONITORING, AND ADHERENCE: A scoping review of international approaches; 2021. https://static1.squarespace.com/static/93856ebbd6827d4bdff1f7e7ae1/t/600ff28547f88841394166bb/1611657863707/939Covid+Isolation+Review+GHGP+20012021.pdf.

60. Marago I, Minen I. Hospital-acquired COVID-19 infection–the magnitude of the problem. Available at SSRN 3622387. 2020.

61. Rhee C, Baker M, Vaidya V, Tucker R, Resnick A, Morris CA, et al. Incidence of nosocomial COVID-19 in patients hospitalized at a large US academic medical center. JAMA network open. 2020; 3(9): e2020498–e2020498. https://doi.org/10.1001/jamanetworkopen.2020.20498 PMID: 32902653