Physics of Life Reviews 51 (2024) 294–327

Contents lists available at ScienceDirect

Physics of Life Reviews

journal homepage: www.elsevier.com/locate/plrev

Investigating and forecasting infectious disease dynamics using epidemiological and molecular surveillance data

Gerardo Chowell [a,c,*], Pavel Skums [b]

[a] Department of Population Health Sciences, School of Public Health, Georgia State University, Atlanta, GA, USA
[b] School of Computing, University of Connecticut, Storrs, CT, USA
[c] Department of Applied Mathematics, Kyung Hee University, Yongin 17104, Korea

ARTICLE INFO

Communicated by: Jose Fernando Fontanari

Keywords:
Epidemic forecasting
Epidemic model
Parameter estimation
Uncertainty quantification
Model validation
Structural identifiability
Practical identifiability
Molecular surveillance
Phylogenetics
Phylodynamics
Molecular epidemiology

ABSTRACT

The integration of viral genomic data into public health surveillance has revolutionized our ability to track and forecast infectious disease dynamics. This review addresses two critical aspects of infectious disease forecasting and monitoring: the methodological workflow for epidemic forecasting and the transformative role of molecular surveillance. We first present a detailed approach for validating epidemic models, emphasizing an iterative workflow that utilizes ordinary differential equation (ODE)-based models to investigate and forecast disease dynamics. We recommend a more structured approach to model validation, systematically addressing key stages such as model calibration, assessment of structural and practical parameter identifiability, and effective uncertainty propagation in forecasts. Furthermore, we underscore the importance of incorporating multiple data streams by applying both simulated and real epidemiological data from the COVID-19 pandemic to produce more reliable forecasts with quantified uncertainty. Additionally, we emphasize the pivotal role of viral genomic data in tracking transmission dynamics and pathogen evolution. By leveraging advanced computational tools such as Bayesian phylogenetics and phylodynamics, researchers can more accurately estimate transmission clusters and reconstruct outbreak histories, thereby improving data-driven modeling and forecasting and informing targeted public health interventions. Finally, we discuss the transformative potential of integrating molecular epidemiology with mathematical modeling to complement and enhance epidemic forecasting and optimize public health strategies.

Abbreviations: COVID-19, Coronavirus disease; US, United State; CDC, Centers for Disease Control and Prevention; ODE, Ordinary Differential Equation; SEIR, Susceptible-Exposed-Infectious-Recovered; GGM, Generalized-Growth Model; GLM, Generalized Logistic Growth Model; DAISY, Differential Algebra for Identifiability of Systems; AMIGO, Advanced Modelling and Identification using Global Optimization; COMBOS, Combinatorial Approach to Structural Identifiability; SIAN, Structural Identifiability Analysis via Numerical methods; AIC, Akaike Information Criterion; BIC, Bayesian Information Criterion; MAE, Mean Absolute Error; MSE, Mean Squared Error; WIS, Weighted Interval Score; MLE, Maximum Likelihood Estimation; NLS, Nonlinear Least Squares; ARE, Average Relative Error; SARS, Severe Acute Respiratory Syndrome; RMSE, Root Mean Square Error; PI, Prediction Interval; ML, Machine Learning; AI, Artificial Intelligence; DL, Deep Learning; SARS-CoV-1, Severe Acute Respiratory Syndrome Coronavirus-1; WHO, World Health Organization; GISAID, Global Initiative on Sharing All Influenza Data; COG-UK, Genomics UK consortium; VOCs, Variants of Concern; VOIs, Variants of Interest; SCOTTI, Structured COalescent Transmission Tree Inference; SOPHIE, SOcial and PHilogenetic Investigation of Epidemics.

* Corresponding author.
E-mail address: gchowell@gsu.edu (G. Chowell).

https://doi.org/10.1016/j.plrev.2024.10.011
Received 22 October 2024; Accepted 23 October 2024
Available online 24 October 2024

1. Introduction

The COVID-19 pandemic has underscored the critical role of mathematical and statistical methods in understanding, forecasting, and mitigating the spread of infectious diseases. Mathematical modeling offers essential tools for public health agencies and researchers worldwide, providing a framework to predict future outbreaks, evaluate intervention strategies, and inform decision-making [1]. These models rely on epidemiological and molecular surveillance data to investigate and track disease transmission dynamics and assess the effectiveness of control measures [2,3]. One specific family of mathematical models based on ordinary differential equations (ODEs) has become central to epidemic analysis [4–6].

For example, epidemic forecasting models have demonstrated their potential to enhance public health responses to various infectious disease outbreaks. They can assist in resource allocation, shape vaccination strategies, and inform policy decisions. During the COVID-19 pandemic, these models provided critical insights into case surges, enabling timely interventions (e.g., [7–10]. The US CDC's FluSight Challenge used models to optimize influenza vaccine distribution [11]. In the West African and Democratic Republic of Congo (DRC) Ebola outbreaks (e.g., [12–14], models predicted disease spread and assessed the effectiveness of interventions. More recently, models have been applied to forecast the spread of mpox, further highlighting their continued importance in epidemic management (e.g., [15–18].

A critical challenge in epidemic modeling is the accurate estimation of parameters, that govern disease transmission and progression in order to generate reliable short-term forecasts of the epidemic's trajectory. Parameters such as transmission rates, intervention effects, and reporting levels are often not directly observable. For this reason, researchers rely on mathematical models, such as ODE-based models (Ordinary Differential Equations), to infer these parameters from available data. Accurate parameter estimation is crucial for the predictive power of these models, as even minor uncertainties in parameter values can lead to significant discrepancies in forecasts.

However, accurate parameter estimation critically depends on both the structural and practical identifiability of the model parameters [19,20]. Indeed, structural and practical identifiability analyses are essential to ensure that model parameters can be uniquely determined and estimated with adequate precision [21]. Structural identifiability guarantees that model parameters can theoretically be identified from perfect data,. whereas practical identifiability examines whether this is achievable with real-world data, which are often noisy or incomplete. Failing to ensure identifiability can lead to unreliable model predictions, undermining the quality of public health decisions. Despite their importance, these analyses are not widely employed in epidemic modeling, underscoring the need for increased attention to these methods in public health modeling efforts. As demonstrated in Fig. 1, there has been a notable increase in epidemic modeling publications that utilize parameter estimation as well as those that mention "identifiability" after 2018, with a sharp rise during the COVID-19 pandemic. However, the proportion of publications that include

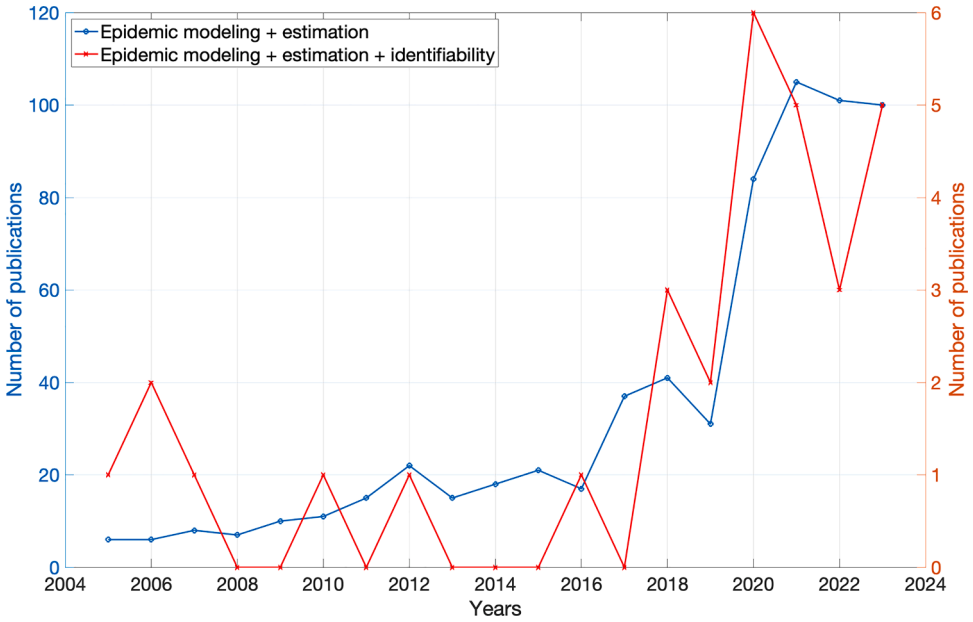


Fig. 1. Trends in publications related to epidemic modeling and estimation and the subset, including identifiability (2004–2023). The blue line shows the number of publications per year, including epidemic modeling and parameter estimation. At the same time, the red line represents the subset of these publications that also include structural or practical identifiability analysis. The y-axis on the left corresponds to the number of epidemic modeling publications. In contrast, the y-axis on the right tracks the number of publications focusing on both estimation and identifiability. The data reveal a notable increase in both categories after 2018, with a sharp rise in publications during the COVID-19 pandemic. However, the proportion of publications that include identifiability is still low, hovering around 5 % during the last few years. Data for this figure was obtained from Web of Science.

identifiability remains low, hovering around 5 % in recent years.

The integration of molecular surveillance into epidemiological monitoring has revolutionized the ability to track and forecast infectious disease dynamics. Traditional methods, such as contact tracing and exposure data collection, while valuable, are labor-intensive and subject to biases. Advances in genomic epidemiology, powered by high-throughput sequencing technologies, now provide a complementary and more precise approach. By analyzing viral genomes in real time, researchers can infer transmission routes, detect emerging variants, and forecast the spread of pathogens with unprecedented accuracy. These methods were instrumental during the SARS-CoV-2 pandemic, enabling the rapid identification of Variants of Concern (VOCs) and improving the reliability of epidemiological forecasts. However, genomic surveillance presents new challenges, including the complexity of genomic data and the need for advanced computational models to process and interpret this information. Despite these hurdles, integrating genomic methods with traditional epidemiological tools offers significant opportunities to enhance the understanding and control of infectious diseases.

This review addresses two critical aspects of infectious disease modeling and monitoring: the methodological workflow for epidemic forecasting and the transformative role of molecular surveillance. First, we examine the iterative methodological workflow necessary for reliably fitting and forecasting epidemic models, encompassing real-time data integration and model formulation or refinement to the generation and evaluation of model-based predictions. In particular, we recommend a structured approach to model validation, systematically addressing key stages such as model calibration, assessment of structural and practical parameter identifiability, and the propagation of uncertainty in forecasts. This structured approach ensures that models are robust, reliable, and capable of supporting evidence-based public health decisions. Practical examples illustrate this workflow using both simulated and real data from the COVID-19 pandemic in Spain. In addition, we explore how the integration of molecular surveillance, particularly genomic epidemiology, has revolutionized infectious disease monitoring, enabling real-time tracking and more precise forecasting of pathogen spread. While this approach presents computational challenges, it significantly enhances traditional methods and offers new opportunities for improved epidemic control.

2. Overview of the workflow for fitting and forecasting with epidemic models

The process of fitting and forecasting epidemic models is an interactive cycle that involves several critical steps. This workflow aims to ensure that the models fit the data well and provide accurate predictions that can be used for decision-making during an epidemic. The schematic in Fig. 2 illustrates the key stages of this iterative process. Initially, real-time data are integrated to continuously refine the model. This step ensures that new information is captured and utilized to update model predictions. The model is then formulated or updated based on the most recent data and understanding of the epidemic. Following the model's formulation, initial conditions and known parameters are determined. Next, the structural identifiability of unknown parameters is assessed. Additional data or model adjustments are required if the parameters are not structurally identifiable. The model is fitted to the data when the unknown parameters are structurally identifiable. Once the model is fitted, the quality of the fit is evaluated. If the fit is unsatisfactory, the workflow loops back to integrating new data or making necessary model adjustments. If the fit is acceptable, the next step is to check the practical identifiability of the estimated parameters. If the parameters are not practically identifiable, the process again requires additional data or model modification. The model is accepted when both the fit quality and the parameters' practical identifiability are confirmed. At this stage, the model is used to generate predictions, and these predictions are evaluated for accuracy and reliability. The iterative nature of this process ensures continuous refinement and improved accuracy over time. The following sections delve into the methods and approaches applicable to each workflow phase. We also illustrate the workflow using simple, practical examples to demonstrate how these methods are applied in real-world scenarios.

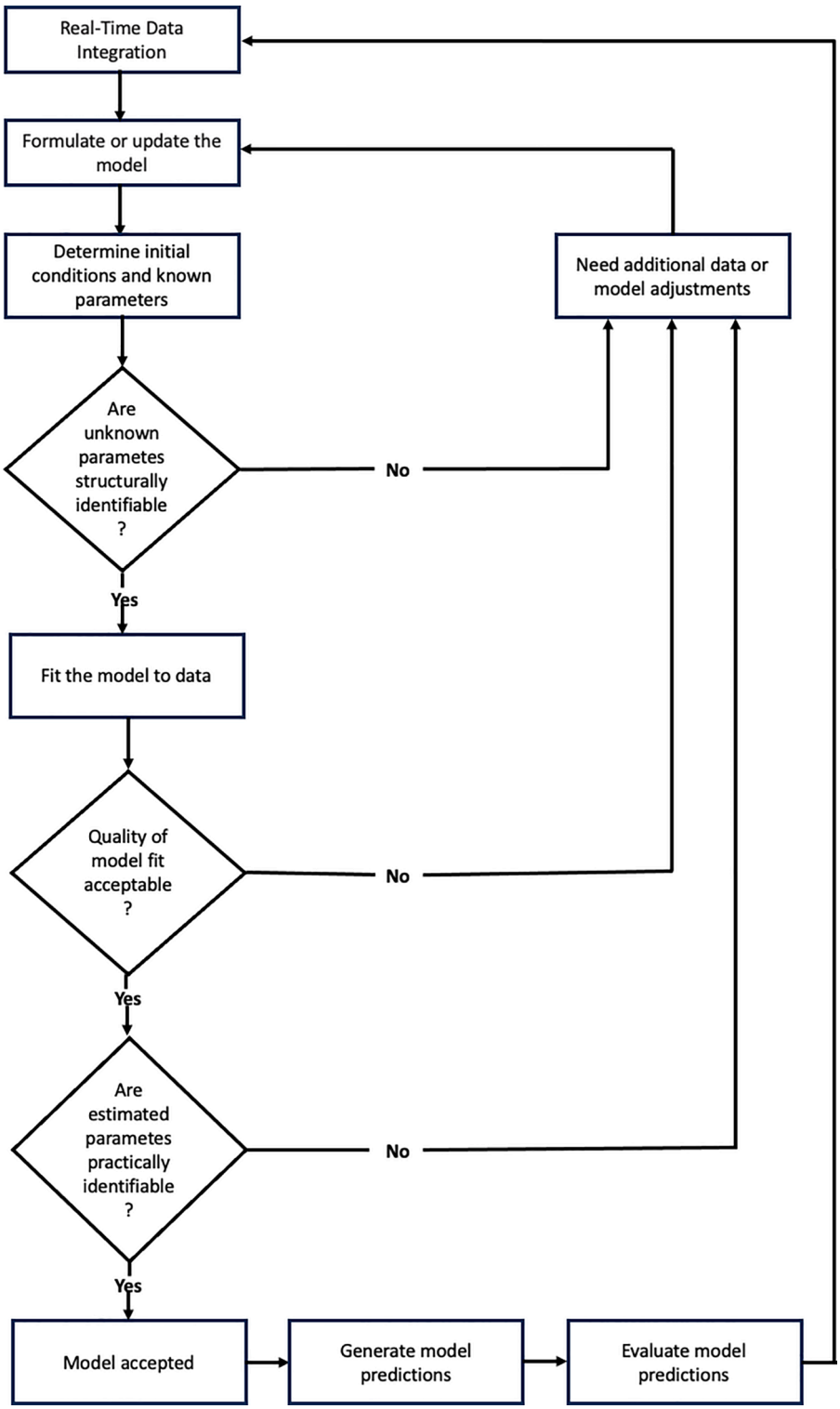
2.1. Real-time data integration

The first phase of the workflow is real-time data integration, which plays a crucial role in ensuring that epidemic models stay current and responsive as new information becomes available or the epidemic situation evolves. In infectious disease outbreaks, the situation can rapidly evolve, with new data on cases, hospitalizations, recoveries, and deaths reported frequently [22]. By integrating this real-time data into the models, public health officials and policymakers can access up-to-date insights vital for making informed decisions [23,24].

Data for real-time integration typically comes from various sources, such as public health surveillance systems, hospital records, diagnostic testing facilities, and even mobile tracking of population movements or infection spread. Incorporating this diverse set of data streams helps ensure a more comprehensive understanding of the epidemic's trajectory [25]. However, these updates must occur rapidly, allowing the model to be adjusted in near real-time to reflect current conditions and enhance situational awareness. This enables the model to remain accurate and responsive to changes in the spread or control of the disease.

One challenge of real-time data integration is ensuring the quality and reliability of incoming data. Data often contain inconsistencies, missing values, or errors that need to be addressed before they can be used [26]. This may involve preprocessing steps such as data cleaning and employing statistical methods or machine learning techniques to fill in gaps or correct anomalies. Ensuring that high-quality data is fed into the model is essential for maintaining the accuracy of forecasts and predictions.

Additionally, automated systems play a critical role in real-time data integration, as they can gather, process, and input data directly into the modeling pipeline, reducing delays and minimizing human error. Through this continuous process of updating, validating, and refining the model with real-time data, predicitons remain relevant and can provide a stronger foundation for responsive public health actions.



(caption on next page)

Fig. 2. Iterative workflow for parameter estimation and forecasting using ODE models. This flow diagram illustrates the critical stages involved in developing and refining epidemic models, emphasizing the iterative nature of the process. The workflow begins with real-time data integration, followed by formulating or updating the model. Initial conditions and known parameters, often derived from early outbreak data, are determined next. The structural identifiability of unknown parameters is then assessed. Additional data or model adjustments are required if parameters are not structurally identifiable. The model is fitted to the observed data if they are structurally identifiable. The next step evaluates the quality of the model fit; if the fit is unsatisfactory, more data or modifications are needed. Once an acceptable fit is achieved, the practical identifiability of the estimated parameters is examined. If parameters are not practically identifiable, further data or adjustments are necessary. The model is accepted only after successfully passing these steps, at which point it can generate and evaluate predictions. This process is essential for real-time epidemic management, ensuring that the model remains accurate and relevant as more data becomes available.

2.2. Formulating the model based on ODEs

The next step in this workflow involves formulating or updating the model in response to real-time data. By continuously refining the model, researchers can better capture the evolving dynamics of the epidemic, enabling more accurate predictions and more effective public health interventions [27]. Epidemic models often take the form of ordinary differential equations, such as the well-known SEIR (susceptible-exposed-infectious-recovered) model [28], which are essential for studying infectious disease transmission dynamics [29]. By describing the rates of change between compartments within a population, these models allow researchers to explore how diseases spread over time. In mathematical epidemiology, such models are critical in assessing intervention strategies, evaluating vaccination effects, estimating key transmission parameters (including the basic reproduction number, R_0 [30,31]), and generating forecasts of the epidemic's trajectory with quantified uncertainty.

ODE models can vary significantly in complexity depending on the number of state variables and parameters. These models are generally categorized into two types: phenomenological models and mechanistic models [32]. Phenomenological models adopt an empirical approach, focusing on reproducing observed trends rather than explaining the underlying biological mechanisms. They are particularly useful for predicting the short-term trajectory of an epidemic, especially during its early stages, when there is considerable uncertainty surrounding the epidemiological characteristics of pathogens [33–35] These models provide a pragmatic solution when detailed biological data are scarce or unavailable allowing for rapid forecasting of potential epidemic growth [36].

In contrast, mechanistic models aim to capture the key transmission as well as behavioral and intervention processes driving the spread of the disease. These models are typically more detailed and complex because they incorporate in-depth information about the interactions between different compartments in the population [37–39]. For instance, the widelyused SEIR model divides the population based on disease status into compartments. More elaborate models, which include compartments for asymptomatic or hospitalized individuals or incorporate vaccination dynamics, offer a more nuanced representation of disease progression. The transitions between these compartments are governed by rates such as the transmission rate (how quickly susceptible individuals become infected), the latent period (the time before exposed individuals become infectious), and the recovery rate (how rapidly infected individuals recover or die).

While phenomenological models focus on fitting observed data, their parameters can sometimes be directly connected to the mechanistic parameters of SIR-type models [40,41]. For instance, in many infectious disease models, including SIR-type models, the growth dynamics often exhibit logistic behavior or closely resemble the logistic model. In these cases, parameters such as the logistic model's growth rate and carrying capacity can be interpreted as key disease transmission parameters like the basic reproduction number (R_0) and recovery rates. This connection arises because the logistic model effectively captures an epidemic's early exponential growth phase, followed by a slowdown as the susceptible population is depleted, which is a common characteristic of epidemic curves. Similarly, the SIS model results in a logistic-like equation that describes the number of infectious individuals over time [42].

An ODE model comprising a system of h ordinary-differential equations is given by:

$$\dot{x_1}(t) = g_1(x_1, x_2, \ldots..x_h, \Theta)$$

$$\dot{x_2}(t) = g_2(x_1, x_2, \ldots..x_h, \Theta)$$

$$\dot{x_h}(t) = g_h(x_1, x_2, \ldots..x_h, \Theta).$$

Above, $\dot{x_i}$ denotes the rate of change of the system state x_i where $i = 1,2,\ldots,h$ and $\Theta = (\theta_1,\theta_2,\ldots,\theta_m)$ is the set of model parameters. Let $f(t,\Theta)$ denote the expected temporal trajectory of the observed state of the system.

In the context of epidemics, the state variables may include observable quantities such as the number of reported cases, deaths, or hospitalizations. These are observed states, while latent states represent variables that are not directly measured but inferred from the mathematical model [43,44]. For example, the number of exposed individuals in a population is often a latent state that can only be indirectly estimated through model fitting.

2.2.1. Phenomenological growth models

Phenomenological growth models are simple yet powerful tools designed to generate real-time short-term forecasts with quantified uncertainty for various growth processes, including epidemic outbreaks. These models focus on empirical data patterns rather than mechanistic explanations, making them particularly useful for characterizing the early stages of growth in both natural and social phenomena. The models discussed here are adept at capturing key epidemic dynamics, such as growth rates, scaling of growth,

doubling times, reproduction numbers, and turning points. They are also effective in predicting critical values, such as the final size of an epidemic at different time horizons, with quantified uncertainty, which makes them highly applicable in real-time decision-making during outbreaks [8,45–51].

These phenomenological models have been frequently used in near real-time to forecast epidemic trajectories, offering practical insights into the evolution of disease outbreaks. Estimating quantities such as the basic reproduction number (R_0) and doubling time provides valuable information that enables public health authorities to better understand and respond to emerging health crises. Their simplicity and ease of implementation make them an attractive option for rapid forecasting and monitoring, particularly in the early phases of an epidemic.

Phenomenological growth models offer flexible tools for capturing the early dynamics of infectious disease outbreaks. Unlike traditional models that typically assume exponential growth, the Generalized-Growth Model (GGM) and Generalized Logistic Growth Model (GLM) incorporate parameters that allow for sub-exponential growth, providing a more accurate representation of varying transmission dynamics across different diseases. This adaptability makes these models particularly useful for forecasting outbreaks with diverse growth patterns. Both the GGM and GLM have been successfully applied to predict the trajectory of epidemics, including Zika, Ebola, and COVID-19 [8,33,47–53].

2.2.2. Generalized-growth model (GGM)

Models commonly used to study the growth pattern of infectious disease outbreaks often assume exponential growth in the absence of control interventions (e.g., compartmental models); however, growth patterns may be slower than exponential for some diseases, depending on the transmission mode and population structure. For instance, diseases like Ebola, which spread via close contact, are expected to exhibit sub-exponential growth patterns due to the constrained nature of population contact structures [54]. The GGM [33] includes a “deceleration of growth” or "scaling of growth" parameter, p (range: [0, 1]), that relaxes the assumption of exponential growth. A value of $p = 0$ represents constant (linear) growth, while a value of $p = 1$ indicates exponential growth. If $0 < p < 1$, the growth pattern is characterized as sub-exponential or polynomial.

The GGM is as follows:

$$\frac{dC(t)}{dt} = C'(t) = rC(t)^p,$$

where the derivative $C'(t)$ describes the incidence curve over time t. The positive parameter r is the growth rate parameter ($r > 0$), and p is the scaling of the growth parameter. For this model, we estimate $\Theta = (r,p)$.

2.2.3. Generalized logistic growth model (GLM)

The GLM extends the GGM by incorporating a final epidemic size parameter and is given by:

$$\frac{dC(t)}{dt} = C'(t) = rC(t)^p\left(1 - \frac{C(t)}{K_0}\right).$$

The growth scaling parameter, p, is also used in the GLM to model a range of early epidemic growth profiles ranging from constant incidence $(p = 0)$, polynomial $(0 < p < 1)$, and exponential growth dynamics $(p = 1)$. When $p = 1$, this model reduces to the logistic growth model. The remaining model parameters are as follows: r is the growth rate, and K_0 is the final cumulative epidemic size (carrying capacity). For this model, $\Theta = (r,p,K_0)$. The GLM has been employed to generate short-term forecasts of Zika, Ebola, and COVID-19 epidemics [2,8,47,51,53].

2.2.4. Mechanistic models

SEIR-type models are widely used compartmental models in epidemic modeling [5,29,55]. They divide the population into compartments representing different epidemiological states of the disease, such as susceptible, exposed, infectious, and recovered. These models capture the natural history of infectious diseases, allowing researchers to simulate transmission dynamics and assess the impact of various interventions. Their flexibility makes them ideal for deriving explicit expressions of the basic reproduction number (R_0) and studying a wide range of infectious diseases [56].

The simplest and most popular mechanistic ODE compartmental model for describing the spread of an infectious agent in a well-mixed population is the well-known SEIR (susceptible-exposed-infectious-removed) model [57]. This model requires four parameters (transmission rate β, the latent period $1/\kappa$, the average infectious period $1/\gamma$, and the population size N) and four state variables that keep track of the number of susceptible, exposed, infectious, and removed individuals over time. Additionally, the models often include an additional state variable to keep track of the number of new infectious individuals over time, frequently used to link the model to time-series data. This model assumes no births or natural deaths in the population. In this model, the infection rate or force of infection is often defined as the product of three quantities: a constant transmission rate (β), the number of susceptible individuals in the population ($S(t)$), and the probability that a susceptible individual encounters an infectious individual $\left(\frac{I(t)}{N}\right)$. Here, $I(t)$ represents the number of infectious individuals in the population at time t, and N is the population size. Exposed individuals (E) become infectious (I) after an average latent period given by $\frac{1}{\kappa}$. Infectious individuals become recovered (R) after an average infectious period given by $\frac{1}{\gamma}$. The model is based on a system of ODEs that keep track of the temporal progression in the number of susceptible (S), exposed (E), infectious (I), and recovered (R) individuals as follows:

$$\begin{cases} \dot{S} = -\beta S(t)\dfrac{I(t)}{N} \\ \dot{E} = \beta S(t)\dfrac{I(t)}{N} - \kappa E(t) \\ \dot{I} = \kappa E(t) - \gamma I(t) \\ \dot{R} = \gamma I(t) \\ \dot{C} = \kappa E(t) \end{cases}$$

In the above system, the auxiliary variable $C(t)$ keeps track of the cumulative number of infectious individuals, and $\dot{C}(t)$ keeps track of the curve of new cases (incidence), which is often used to link the model to time series data. If $f(t,\Theta)$ denotes the temporal trajectory of the observed state of the system, then $f(t,\Theta)$ will correspond to $\dot{C}(t)$ in the SEIR model above.

In a completely susceptible population, e.g., $S(0) \approx N$, the number of infectious individuals grows following an exponential function during the early epidemic growth phase, e.g., $I(t) \approx I_0 e^{(\beta-\gamma)t}$. Moreover, the basic reproduction number (R_0) quantifies the average number of secondary cases generated per primary case during the initial transmission phase. This parameter is a function of several parameters of the epidemic model, including the transmission rates and infectious periods that contribute to new infections. R_0 often serves as a threshold parameter for the SEIR-type compartmental models. If $R_0 > 1$ then an epidemic is expected to occur whereas values of $R_0 < 1$ cannot sustain disease transmission. For this simple SEIR model, R_0 is simply given by the product of the mean transmission rate (β) and the mean infectious period ($\frac{1}{\gamma}$) as follows: $R_0 = \frac{\beta}{\gamma}$.

2.2.5. SEIR model with reported and unreported cases

This extended SEIR model incorporates a reporting proportion parameter (ρ) to account for underreporting, distinguishing between reported (I_r) and unreported (I_u) infectious individuals. The model also includes an equation to track the cumulative number of reported cases over time, $C(t)$, which provides insight into the total number of individuals detected and reported as infectious. This model is governed by the following system of differential equations:

$$\begin{aligned} \frac{dS}{dt} &= -\beta \frac{S\,(I_r + I_u)}{N} \\ \frac{dE}{dt} &= \beta \frac{S\,(I_r + I_u)}{N} - \kappa E \\ \frac{dI_r}{dt} &= \rho\,\kappa\,E - \gamma\,I_r \\ \frac{dI_u}{dt} &= (1-\rho)\,\kappa\,E - \gamma\,I_u \\ \frac{dR}{dt} &= \gamma\,(I_r + I_u) \\ \frac{dC}{dt} &= \rho\,\kappa\,E \end{aligned}$$

2.3. Determining initial conditions and known parameters

Once the model is formulated, the initial conditions for each compartment and any known parameters must be specified. Initial conditions represent the system's state at the epidemic's beginning (e.g., the number of infected and susceptible individuals). Known parameters, such as population size and other well-established epidemiological parameters, can be obtained from previous studies or early outbreak investigations. Accurate specification of these conditions is crucial because they influence the model's behavior and ability to capture the epidemic dynamics [58]. For instance, some models focus on the initial phase of an outbreak, where the population is assumed to be large, and the depletion of susceptible individuals is negligible. In this phase, each infected individual typically produces a new generation of secondary infections. This is crucial for calculating early epidemic dynamics, including R_0 [2,57].

Whether the initial conditions are known or unknown has significant implications for the structural identifiability of other parameters in the model, as discussed in the next section. If the initial conditions are known, the process of parameter identifiability is simplified because the model has fewer degrees of uncertainty. This allows for a more straightforward estimation of other model parameters, as the starting state of the epidemic is fixed. Conversely, when the initial conditions are unknown, they become part of the parameter estimation problem, increasing the model's complexity and making it more challenging to uniquely identify the remaining parameters. To address this challenge, using prior knowledge from earlier outbreaks can help refine initial condition estimates during calibration. In such cases, the structural identifiability analysis must account for the joint uncertainty of both the initial conditions and the parameters of interest.

2.4. Structural identifiability of parameters

Before calibrating the model with real data, it is crucial to determine whether the unknown parameters are structurally identifiable [59–62]. This step in the workflow involves assessing whether the model's structure allows for the unique estimation of parameters based on the available data. If parameters are not identifiable, multiple combinations of parameter values could explain the data equally well, introducing uncertainty into the model's predictions and potentially compromising decision-making [63]. A lack of identifiability often leads to ambiguous interpretations of model outputs, since different parameter sets may yield the same outcomes, which can mask the true dynamics of the epidemic [20]. This undermines the ability to derive clear insights, for instance, about transmission rates or intervention efficacy, and can lead to poor policy recommendations.

Structural identifiability problems frequently arise from high correlations among parameters and redundant parameters [42]. As the number of parameters in the model increases, the information from the data is distributed across the parameters, increasing the likelihood of encountering structural identifiability issues. This creates a situation where the model becomes overparameterized relative to the available data, diluting the informative value of each parameter and leading to non-unique or unstable parameter estimates.

In the context of epidemic models, a structurally identifiable model ensures that key parameters—such as transmission rates, recovery rates, and other epidemiological factors—can be uniquely estimated from available data, such as case counts or mortality reports.

The structural identifiability problem mathematically determines whether the model parameters can be uniquely recovered from the model's outputs y(t). A model is structurally identifiable if the parameters (p) to y(t) mapping is injective. This means that different parameter values must produce distinct outputs Formally, this requires showing that:

$$y(t,p_1) = y(t,\ p_2) \text{ implies } p_1 = p_2.$$

If the model fails this condition, it is structurally non-identifiable, implying that multiple combinations of parameters can produce the same output data, thereby leading to parameter uncertainty and ambiguity in model predictions. This issue is especially problematic in real-time epidemic modeling, where decision-makers rely on clear parameter estimates to shape interventions, such as determining the intensity of social distancing or vaccination campaigns.

Several methods have been developed for structural identifiability analysis, each offering distinct advantages depending on the complexity of the model. These methods include approaches that expand the model's equations using series expansions, such as the Taylor series method [64], which analyzes the identifiability through the series expansion of the model's equations. Other methods are the generating series method [65], which uses formal power series for systematic analysis, the similarity transformation approach [66], which transforms the system to new coordinates, the direct test method [67], which tests the model output equations through differentiation, and the differential algebra method [68,69], which uses symbolic manipulations to assess identifiability.

Several software tools, such as DAISY (Differential Algebra for Identifiability of Systems) [68], AMIGO [70], COMBOS (Combinatorial Approach to Structural Identifiability) [71], SIAN (Structural Identifiability Analysis via Numerical methods) [72] and *structuralidentifiability.jl* [73] are available for investigating the structural identifiability of model parameters. One of the most widely used techniques is the Differential Algebra Method, which expresses the model's equations in terms of observable variables and parameters, helping to identify parameter correlations that could hinder identifiability [68]. Tools like DAISY automate much of this analysis using the computer [74]. In addition to these traditional tools, new. computational tools such as structuralidentifiability.jl have emerged to provide more efficient and scalable solutions [73]. This Julia-based package leverages symbolic computations to perform structural identifiability analysis and tends to work well with more complex models.

When structural identifiability analysis reveals non-identifiable parameters, several approaches can be employed to resolve these issues. One significant consequence of working with non-identifiable parameters is that it can lead to unstable or unreliable forecasts, and hinder the ability to attribute changes in epidemic dynamics to specific interventions, as different parameter combinations may explain the same trends. One common solution to identifiability issues is to re-parameterize the model [42]. In cases where two parameters are indistinguishable, re-parameterizing allows the model to express the same relationship between parameters more easily. Another effective method is fixing parameters, where known values for certain parameters, such as population size, are used to resolve correlations and make the remaining parameters identifiable [75,76]. This strategy is especially useful when some parameters are already well understood, allowing the focus to shift to estimating more uncertain variables. Incorporating additional observed states is another approach that can enhance identifiability. For example, collecting additional data on other states of the model can help identify all relevant parameters. In this case, additional information from different epidemic stages—such as hospitalization or recovery data—can help resolve parameter correlations. Additionally, using prior knowledge to fix correlated parameters is a key strategy for managing non-identifiability issues [77]. Finally, simplifying model structure by reducing the number of compartments can sometimes resolve identifiability problems by making the model less complex and more manageable. These strategies are crucial for ensuring that parameters in epidemic models are uniquely and accurately estimated from the observed states. However, this simplification must be done carefully to avoid losing critical epidemiological detail.

It is important to note that structural identifiability is a theoretical property of the model based on the assumption of perfect, noise-free data. However, data are often noisy, sparse, and incomplete in practice. Therefore, even if a model is structurally identifiable, its parameters may not be practically identifiable. The inability to practically estimate parameters in real-world contexts can result in misleading confidence in the model's outputs, further complicating intervention planning. Practical identifiability, which is discussed in the next section, considers real-world factors such as data quantity, measurement noise, and sampling frequency. These factors can

significantly affect the precision with which parameters can be estimated in real-world scenarios.

Modelers need tools that precisely calibrate ODE models using time series data to perform practical identifiability analyses. The following section discusses the available tools for both frequentist and Bayesian inference frameworks, which are essential for fitting models to real-world data.

2.5. Fitting the model to data and evaluating the goodness of fit

The next step in the workflow involves fitting the model to observed data, such as case counts, hospitalizations, or deaths, to evaluate how well the model captures observed trends and simultaneously assess the practical identifiability of its parameters. This process requires estimating unknown parameters using either frequentist methods, such as maximum likelihood estimation, or Bayesian inference. The primary objective is to find parameter values that minimize the discrepancy between the model's predictions and the actual data, ensuring a robust fit. The quality of the fit is evaluated using goodness-of-fit measures, such as the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC), as well as performance metrics like mean absolute error (MAE), mean squared error (MSE), coverage of the 95 % prediction interval, and the weighted interval score (WIS). This section reviews contemporary tools that facilitate model fitting, assess fit quality, evaluate practical identifiability, and generate forecasts with quantified uncertainty.

One of the simplest approaches to fitting models to data involves minimizing the distance between observed data and model predictions, often through techniques like nonlinear least squares [78–80]. While effective for capturing the average trajectory that fits the data, this method assumes a constant variance in the errors, which may not yield optimal parameter estimates when the variance is not uniform (heteroscedasticity). In such scenarios, maximum likelihood estimation provides a more flexible approach by allowing for error structures that reflect realistic uncertainty around the model fit, enabling well-calibrated prediction intervals [81]. Following this paradigm, nonlinear optimization algorithms are utilized to maximize the likelihood function, which expresses how well the model fits the data.

In the broader landscape of tools available for fitting models to data, both the *fitode* package [81] and *QuantDiffForecast* toolbox [82] offer parameter estimation and model fitting capabilities for ordinary differential equation (ODE) models, but they cater to different user needs. *fitode*, an R-based package, is optimized for simplicity and user accessibility, allowing researchers to fit ODE models using Maximum Likelihood Estimation (MLE). It offers flexibility by supporting various error structures, such as normal and negative binomial distributions. This makes *fitode* well-suited for applications that require rapid model fitting and basic uncertainty assessment without significant computational overhead.

In contrast, *QuantDiffForecast* is a MATLAB-based toolbox designed to facilitate parameter estimation and short-term forecasting with a strong emphasis on quantified uncertainty. It uses parametric bootstrapping to characterize probabilistic predictions and construct confidence prediction intervals. Additionally, it supports the use of multiple initial seeds (e.g., the Multistart approach) to enhance the likelihood of the optimization algorithm converging to a global optimum rather than getting trapped in a local one. Its user-friendly interface and detailed tutorials make it accessible to both students and professionals working with dynamic systems. The toolbox streamlines key processes such as parameter estimation, model fitting, and uncertainty quantification. In addition to supporting error structures like Poisson and negative binomial, QuantDiffForecast provides calibration and forecasting performance metrics such as MAE and WIS [83], making it highly suitable for real-time forecasting and policy evaluation. QuantDiffForecast's flexibility is enhanced through the inclusion of options files, which allow users to customize model settings, run batch simulations, and replicate configurations easily, streamlining complex modeling workflows. It supports the definition of ODE models, accommodating both phenomenological and mechanistic models, and offers flexibility in parameter estimation through nonlinear least squares (NLS) and MLE. Notably, the toolbox addresses real-world data complexities by supporting various error structures, such as Poisson and negative binomial distributions, making it well-suited for handling complex datasets [84]. Moreover, the toolbox is equipped to provide short-term forecasts with quantified uncertainty, offering users metrics such as MAE, MSE, the coverage of 95 % prediction intervals, and the WIS to assess forecast performance. By integrating these features, *QuantDiffForecast* offers a comprehensive platform for model fitting and forecasting model trajectories.

Complementing *QuantDiffForecast, BayesianFitForecast* is an R-based toolbox that simplifies Bayesian parameter estimation and forecasting for ODE models [85]. While *QuantDiffForecast* focuses on frequentist approaches such as NLS and MLE, *BayesianFitForecast* allows researchers to incorporate prior knowledge into the model-fitting process, offering an alternative framework for parameter estimation through Bayesian inference [86,87]. This toolbox is particularly useful for researchers using ODEs to model dynamic systems. By automatically generating Stan code [88], *BayesianFitForecast* reduces the need for extensive programming expertise, providing a user-friendly way to estimate parameters using advanced optimization techniques. Like *QuantDiffForecast, BayesianFitForecast* also supports flexible error structures, allowing users to choose between Poisson, negative binomial, or normal distributions, making it suitable for handling overdispersed data.

For forecasting, *BayesianFitForecast* generates real-time predictions with uncertainty quantification, offering similar performance metrics as *QuantDiffForecast*. This comprehensive approach allows modelers to assess the quality of both parameter estimates and forecasts.

It is worth mentioning that alternative tools exist for fitting stochastic models, such as the *pomp* package in R [89]. This package was designed for fitting models that incorporate process error, such as demographic stochasticity or random fluctuations in transmission rates, which affect both current and future states of a trajectory. However, our focus here is on methods based on deterministic ODE models while accounting for observation error. It is important to note that having a model that fits the data well is insufficient; ensuring that the estimated parameters are practically identifiable is equally critical. These aspects of practical parameter identifiability are

addressed in the next section.

2.6. Practical identifiability of parameters

Once structural identifiability of the parameters is established and the model fits the data well, the next step in the workflow involves is to evaluate the practical identifiability of the model parameters using real data. While structural identifiability ensures that model parameters can, in theory, be uniquely determined from ideal, noise-free data, practical identifiability assesses whether parameters can be reliably estimated from real-world data under conditions of measurement noise and irregular data collection [20,21,76,90,91]. Data are often noisy, incomplete, and collected at irregular intervals, which can significantly hinder parameter estimation—even when the model is structurally identifiable [77,92]. A lack of practical identifiability can be detected by examining the confidence intervals of the parameters after fitting the model. If a parameter is practically unidentifiable, its confidence interval will be wide, indicating that the data provide insufficient information to estimate the parameter precisely. In contrast, well-identified parameters typically have narrow, well-defined confidence intervals. Several metrics can be used to assess whether a parameter is practically identifiable, including a profile likelihood that does not exceed a critical threshold for large increases or decreases in the parameter [21], a large relative standard error of the parameter (e.g., $>30\,\%$), a high coefficient of variation (e.g., greater than 1), and a small near-zero eigenvalue of the Fisher Information Matrix.

Several factors influence the practical identifiability of model parameters. Foremost among them are the quality and quantity of data are critical factors, as high-quality data with minimal noise and frequent measurements improve the likelihood of obtaining precise parameter estimates [76,243]. Conversely, noisy or sparse data can lead to ambiguous results and wide confidence intervals, making distinguishing between different parameter values difficult. Additionally, an insuficient number of observations relative to the number of parameters can result in overfitting, where the model captures noise instead of true dynamics, leading to unreliable predictions [93]. Measurement noise further complicates parameter estimation, as real-world data often exhibit non-constant variance or overdispersion, which requires more specific techniques to characterize the relationship between model variables, parameters, and time-series data [94]. Even structurally identifiable models may struggle with noisy data, resulting in poor parameter estimates. Sampling frequency is another important consideration, especially in epidemic modeling. Infrequent data collection may fail to capture rapid changes in disease dynamics, whereas more frequent sampling can improve the precision of parameter estimates. Lastly, model complexity affects practical identifiability. More complex models with many parameters typically require more data for precise estimation. Simpler models, while easier to fit, may lack the flexibility to capture important dynamics, highlighting the challenge of balancing model complexity with data availability.

Several techniques are commonly used to assess the practical identifiability of model parameters. Profile likelihood is one of the most widely used methods, where one parameter is varied over a range of values, and for each value, the likelihood is maximized with respect to the other parameters [95,96]. A flat likelihood over a wide range indicates poor identifiability, while a sharp peak suggests that the parameter is well-identified. Simulation-based approaches involve generating synthetic data with known parameter values and estimating these parameters from the simulated data with different levels of noise and sampling frequencies to quantify the average relative error (ARE) between the original and re-estimated parameters [78,91]. This helps assess how well the model recovers the true parameters under different levels of noise and sampling frequencies and can be used to determine the minimum amount of data needed for reliable estimates [97,98].

To improve identifiability, when parameters are poorly identifiable in practice, several strategies can be employed. One of the most straightforward approaches is to collect more data, either by increasing the frequency of data collection or by adding additional variables. For instance, gathering data on both infections and hospitalizations in epidemic modeling can provide more information for estimating key parameters like transmission and recovery rates [99]. Another strategy is to simplify the model by reducing the number of parameters, which may involve merging compartments or fixing certain parameters based on prior knowledge. Simplified models often yield more stable parameter estimates, particularly when data are limited.

2.7. Model validation and forecast generation

Once the model has been validated and is well-calibrated, meaning the parameters are practically identifiable with reasonable uncertainty and the model fits the data well, it can be confidently used to generate forecasts that provide valuable insights into the future trajectory of the epidemic. These forecasts can predict the expected number of cases or hospitalizations over time. Incorporating both real-time data and intervention measures is crucial for improving the accuracy of these forecasts, as Zhang et al. [99] emphasize the need to model time-dependent parameters reflecting changing interventions or behaviors. Continuous validation of the model's predictions by comparing them with newly collected data as the epidemic progresses. As new information becomes available or the epidemic dynamics shift—due to factors such as interventions or behavioral changes—it may be necessary to adjust the model to ensure it remains accurate and relevant for ongoing decision-making.

In both frequentist and Bayesian frameworks, it is essential to quantify forecast uncertainty. Based on the frequentist approach described in [82] we can make h time units ahead forecasts using the estimate $f(t+h,\widehat{\Theta})$. The uncertainty of the forecasted value can be obtained using the model fits of the bootstrap samples using the frequentist approach. Let

$$f(t+h,\widehat{\Theta}_1),\ f(t+h,\widehat{\Theta}_2),\ldots,f(t+h,\widehat{\Theta}_B)$$

denote the forecasted value of the current state of the system propagated h time units ahead, where $\widehat{\Theta}_b$ denotes the estimation of

parameter set Θ from the b_{th} bootstrap sample. We can use these values to calculate the bootstrap variance to measure the uncertainty of the forecasts and use the 2.5 % and 97.5 % percentiles to construct the 95 % prediction intervals (PI) with the assumed error structure. In a Bayesian estimation framework, forecasts can be derived similarly by sampling from the posterior distributions of the parameter values [12,100].

2.8. Iteration and model refinement

The forecasting process is inherently iterative, as models evolve over time to incorporate new features of the epidemic process and additional data (Fig. 2). When parameters are not structurally or practically identifiable or the model does not fit the data well, iteration and refinement, as outlined in the workflow diagram, are essential. This process may also include sensitivity analysis to assess how different parameters impact model outcomes. These refinements may include adjusting the model structure, re-estimating parameters, or integrating new data to address identifiability issues and enhance model accuracy. This cyclical process continues until the model produces reliable forecasts, empowering public health officials to make informed decisions. To effectively guide response strategies, it is critical that the model remains adaptable to incoming data and responsive to changes in epidemic dynamics while incorporating uncertainties in projections.

3. A practical example using simulated data with multiple data streams

During epidemic emergencies such as the 2002–2003 SARS outbreaks, the 2009 A/H 1N1 influenza pandemic, the 2014–2016 Ebola epidemic in West Africa, and the recent COVID-19 pandemic, modeling studies have highlighted significant challenges in predicting the trajectory of outbreaks, particularly in forecasting the peak of an epidemic [27,101–108]. These difficulties often arise from models that rely on a single time series of reported cases, limiting their ability to generate reliable predictions of the outbreak's progression. The use of a single data stream, influenced by inconsistencies in reporting practices, testing rates, and varying public health interventions, introduces substantial variability in data quality and availability. Moreover, unpredictable changes in population behavior in response to the evolving epidemic further complicate predictions.

To address these limitations, it is crucial to develop modeling approaches that rely on multiple data streams—such as reported cases, deaths, recoveries, and hospital occupancy—to improve parameter identifiability and the precision of forecasts. Adding more observed states enhances practical identifiability by providing complementary information that allows for more precise and unique estimation of model parameters [109]. In the following example, we investigate the impact of varying the number of observed states and the length of the calibration period on the accuracy of SEIR model forecasts. By examining 30-day ahead forecasts and evaluating performance metrics, we aim to demonstrate how adding additional observed states and extended calibration periods can improve the reliability of model predictions. This analysis provides valuable insights for enhancing the precision of forecasts during ongoing and future epidemic responses.

3.1. Model and data

We used the SEIR compartmental model to evaluate the impact of calibration data and observed states on forecast accuracy and practical identifiability of parameters. The SEIR model was fitted to simulated data streams representing different numbers of observed states (1, 2, or 3 states), which included newly reported cases (dC/dt), the number of infectious individuals (dI/dt), and recoveries (dR/dt). Each additional observed state introduces new dimensions of information that help resolve uncertainties in parameter estimates,

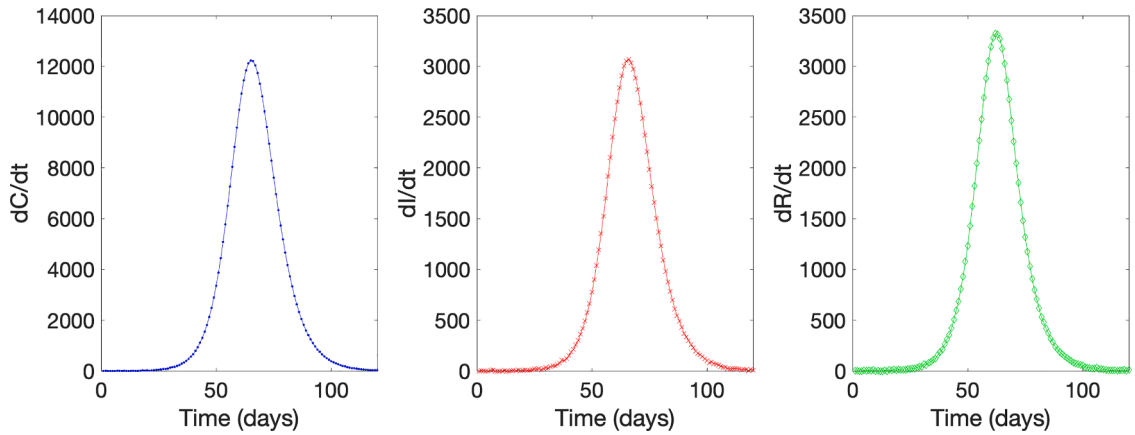


Fig. 3. The simulated time series from the deterministic SEIR model with constant parameters. We added normally distributed noise with a standard deviation of 5 to the simulated curves of the number of new cases (dC/dt), number of infectious individuals $I(t)$, and the number of new recovered individuals (dR/dt).

improving the practical identifiability of the model. We assessed model performance over different calibration periods (ranging from 30 to 60 days in 5-day increments) by varying the amount of data used for calibration, examining periods ranging from shorter durations to more extended calibration windows.

We generated synthetic data from the deterministic SEIR model with constant parameters. We added normally distributed noise with a standard deviation of 5 to the simulated curves of the number of new cases (dC/dt), number of infectious individuals $I(t)$, and the number of new recovered individuals (dR/dt). The simulated time series covering 100 days is shown in Fig. 3. For each combination of calibration period and number of observed states, the model parameters were estimated by fitting the SEIR model to the available data through maximum likelihood estimation using the previously discussed *QuantDiffForecast* toolbox [82].

The parameter estimates (mean values and 95 % confidence intervals) were computed, allowing us to assess the influence of additional observed states and longer calibration periods on the precision of the parameter estimates. With more observed states, the model benefits from diverse data sources that constrain parameter estimates more effectively, reducing uncertainty and improving practical identifiability. For example, while reported cases alone might be insufficient to identify certain transmission dynamics, adding data on recoveries or hospital occupancy provides additional constraints that improve parameter estimation. The confidence intervals for the parameter estimates were calculated using bootstrapping methods to account for estimation variability [78].

3.2. Forecasting and performance metrics

We generated 30-day ahead forecasts by calibrating the SEIR model using different calibration periods and numbers of observed states. Forecasts were computed iteratively by incorporating new data as they become available, to assess the model's ability to predict future outcomes based on the number of states used and the length of the calibration period. Forecast uncertainty was quantified by constructing prediction intervals (95 % confidence) around the forecasted values. The performance of these forecasts was evaluated by comparing predicted values to actual observed data over the 30-day forecast horizon. To assess the accuracy of the forecasts, we calculated performance metrics such as the MAE and the MSE, the coverage of the 95 % prediction interval and the weighted interval

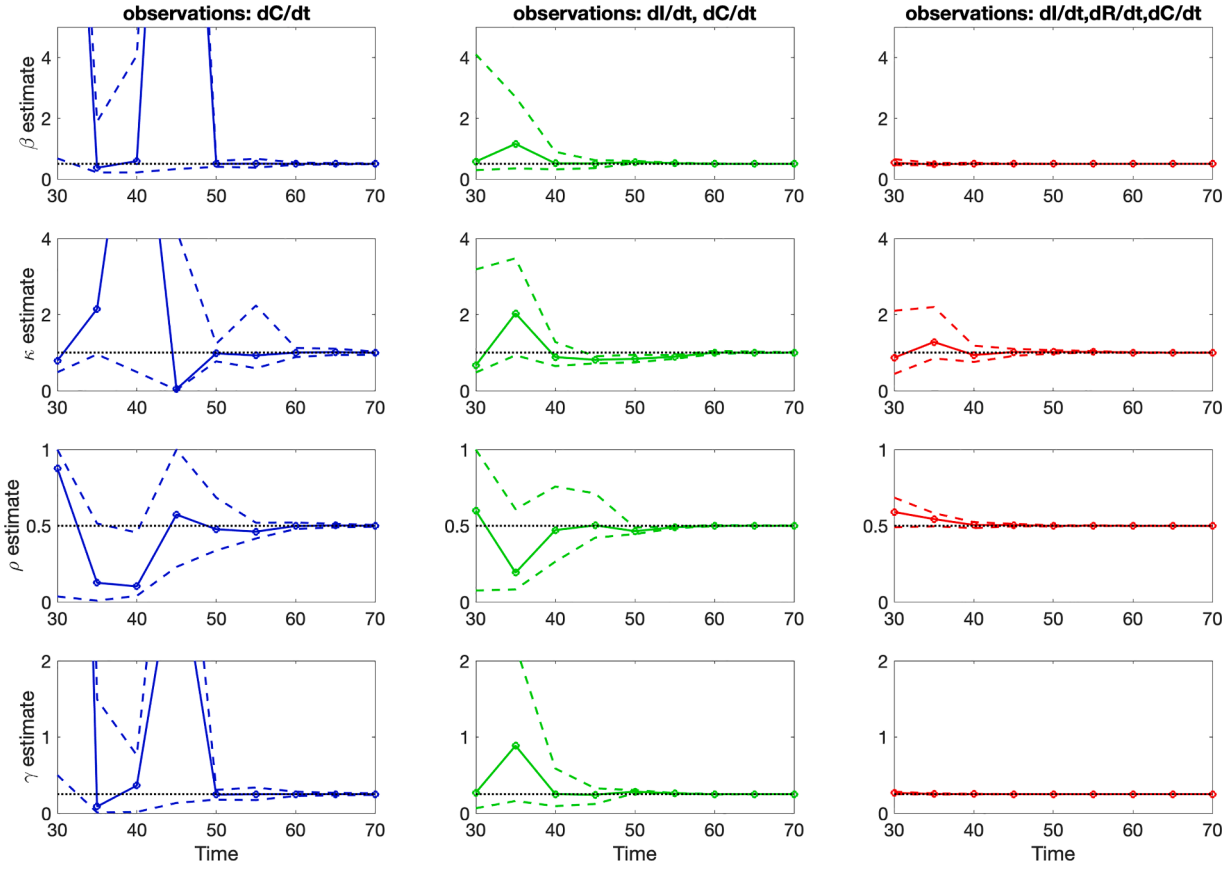


Fig. 4. Parameter estimates (solid line represents the mean and dashed lines show the 95 % confidence intervals) as a function of the amount of data used for model calibration. The estimates for the SEIR model are shown using 1, 2, or 3 observed states. The figure demonstrates that including additional observed states and longer calibration periods significantly enhances the precision of parameter estimates. As more data are incorporated, the variability in the estimates decreases, with longer calibration periods generally producing narrower confidence intervals, indicating increased confidence in the parameter estimates.

score. These metrics were averaged across different calibration periods and analyzed as a function of the number of observed states used in the model. This analysis enabled us to quantify how including more data streams and longer calibration windows affects the model's predictive performance.

The parameter estimates derived from the SEIR model, using different numbers of observed states (1, 2, or 3) and varying calibration periods, demonstrate notable trends in precision and variability. As shown in Fig. 4, adding additional observed states consistently improved the accuracy of parameter estimates, with narrower 95% confidence intervals observed when more states were used for calibration. This improvement in parameter precision is directly linked to enhanced practical identifiability. Each observed state provides distinct data that help resolve ambiguities in the parameter space, allowing the model to converge on more accurate estimates. Additionally, longer calibration periods resulted in further reductions in parameter variability, suggesting that more extensive data collection enhances the reliability of the estimated parameters. Models using only one observed state had broader confidence intervals, indicating greater parameter estimation uncertainty than models incorporating two or three observed states.

The 30-day ahead forecasts generated by the SEIR model, shown in Fig. 5, reveal the impact of both the calibration period length and the number of observed states on forecast reliability. Forecasts based on longer calibration periods and more observed states (two or three) produced narrower prediction intervals, reflecting greater confidence in the forecasted epidemic trajectory. The improvement in forecast reliability stems from the improved practical identifiability of parameters, as additional observed states help the model better capture the underlying dynamics of disease transmission. In contrast, forecasts using shorter calibration periods or only a single observed state exhibited wider intervals, indicating higher uncertainty. Notably, the forecasts generated from models calibrated with three observed states over an extended period aligned more closely with the actual observed data, suggesting that adding additional data streams enhances forecast precision by improving practical identifiability.

The performance metrics for 30-day forecasts as a function of the calibration period and the number of observed states are displayed in Fig. 6, and the average performance metrics are presented in Fig. 7. The corresponding performance metrics demonstrate a clear improvement in forecast accuracy as more observed states were incorporated into the calibration process. Forecasts calibrated with three observed states consistently outperformed those based on one or two states, with lower error values across all calibration periods. Additionally, forecasts generated from longer calibration periods showed reduced error rates, emphasizing the importance of

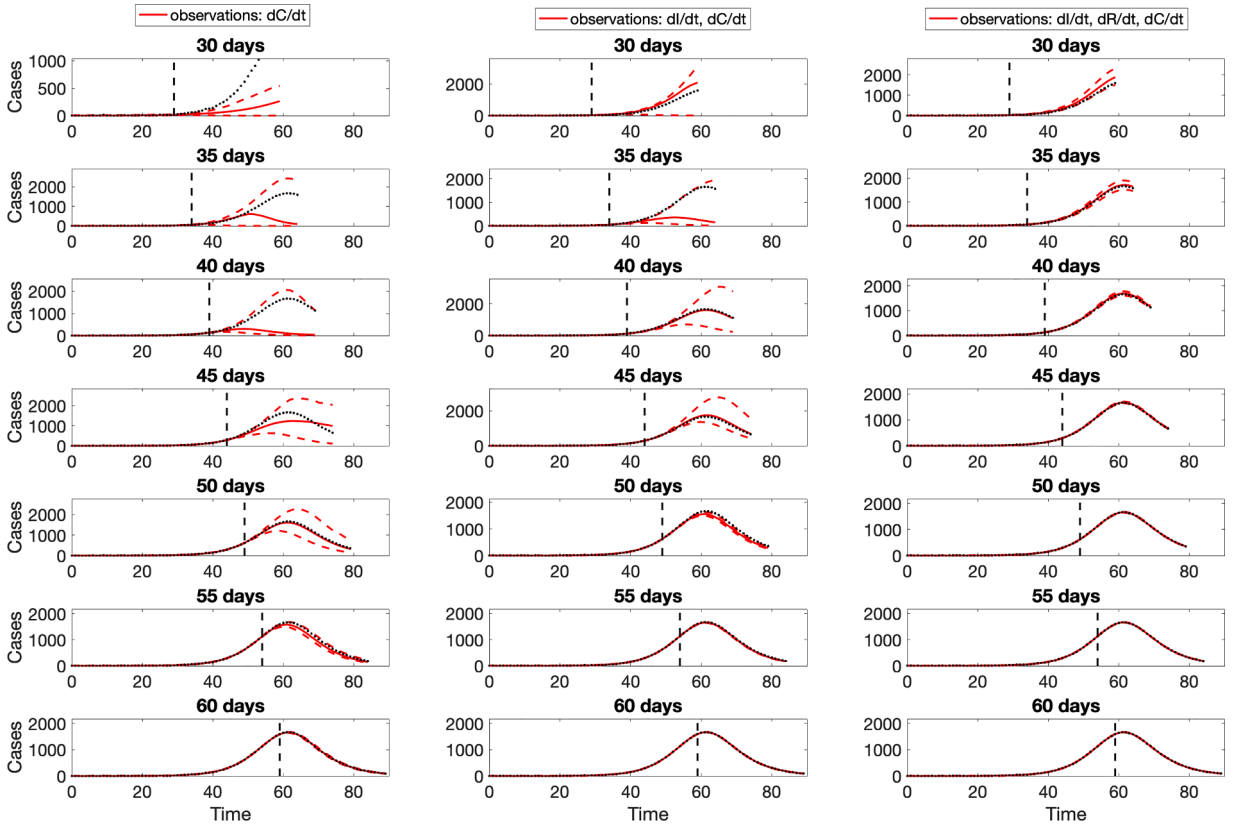


Fig. 5. 30-day ahead forecasts generated by the SEIR model as a function of the calibration period and the number of observed states used for calibration. The figure shows how the calibration period's length and the number of observed states directly impact forecast accuracy and uncertainty. Longer calibration periods and multiple observed states result in more reliable forecasts with narrower prediction intervals. This indicates increased confidence in the model's projections, as adding additional data streams reduces forecast uncertainty. Black dots correspond to the simulated data points. The mean fit (solid red line) and 95 % prediction interval (dashed lines) are also shown. The vertical line separates the calibration period (left) from the forecasting period (right).

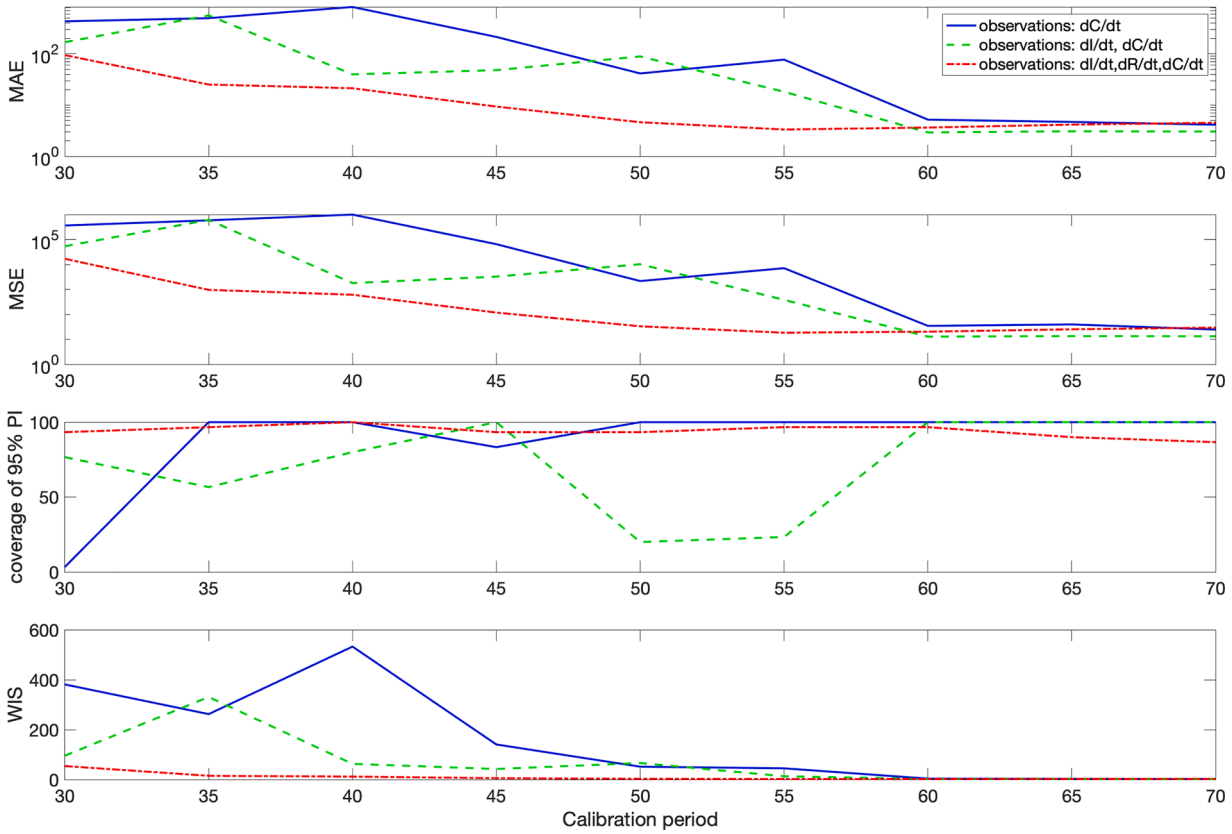


Fig. 6. Performance metrics for 30-day ahead forecasts from the SEIR model as a function of increasingly longer calibration periods and the number of observed states used for model calibration. The figure highlights clear improvements in forecast accuracy as both the calibration period and the number of observed states increase. The performance metrics illustrate how model accuracy improves with extended calibration periods and the inclusion of additional observed states, emphasizing the importance of using multiple data streams to enhance model performance.

using extended data for model calibration. These findings highlight the benefits of incorporating multiple data streams and longer calibration windows to improve forecast performance in epidemic modeling.

In summary, this example highlights the importance of using multiple observed states and extended calibration periods to improve forecasts' practical identifiability and reliability in epidemic modeling. The results demonstrate that models relying on a single observed state and shorter calibration periods tend to produce less reliable parameter estimates and forecasts, with broader confidence and prediction intervals reflecting greater uncertainty. In contrast, adding additional observed states, such as reported cases, hospitalizations, and recoveries, significantly enhances the precision of parameter estimates and reduces forecast uncertainty, particularly when combined with longer calibration periods. Increased practical identifiability using multiple observed states ensures that model parameters are uniquely identifiable, leading to more accurate and reliable forecasts. By increasing practical identifiability through the use of multiple observed states, models can provide more reliable predictions, ultimately supporting more effective public health interventions during epidemic emergencies.

4. A practical example from the early wave of COVID-19 in Spain

The first wave of COVID-19 in Spain began in late February 2020, with rapid growth in cases by early March, particularly in Madrid and the Basque Country. On March 14, 2020, the government declared a state of emergency, implementing some of Europe's strictest lockdowns, which confined citizens to their homes and closed schools, non-essential businesses, and public spaces. The epidemic curve surged, with the peak occurring in late March and early April. Hospitals, especially in Madrid and Catalonia, faced overwhelming pressure as daily deaths exceeded 900. The early intervention measures, including extended lockdowns, helped slow the spread of the virus. By early May, Spain began a phased de-escalation process, gradually lifting restrictions as cases declined. The first wave officially ended when the emergency was lifted on June 21, 2020, although mask mandates and social distancing remained in place [110].

4.1. Investigating the early growth dynamics

When an infectious disease outbreak spreads within a population, it is essential to examine the early growth phase using simple

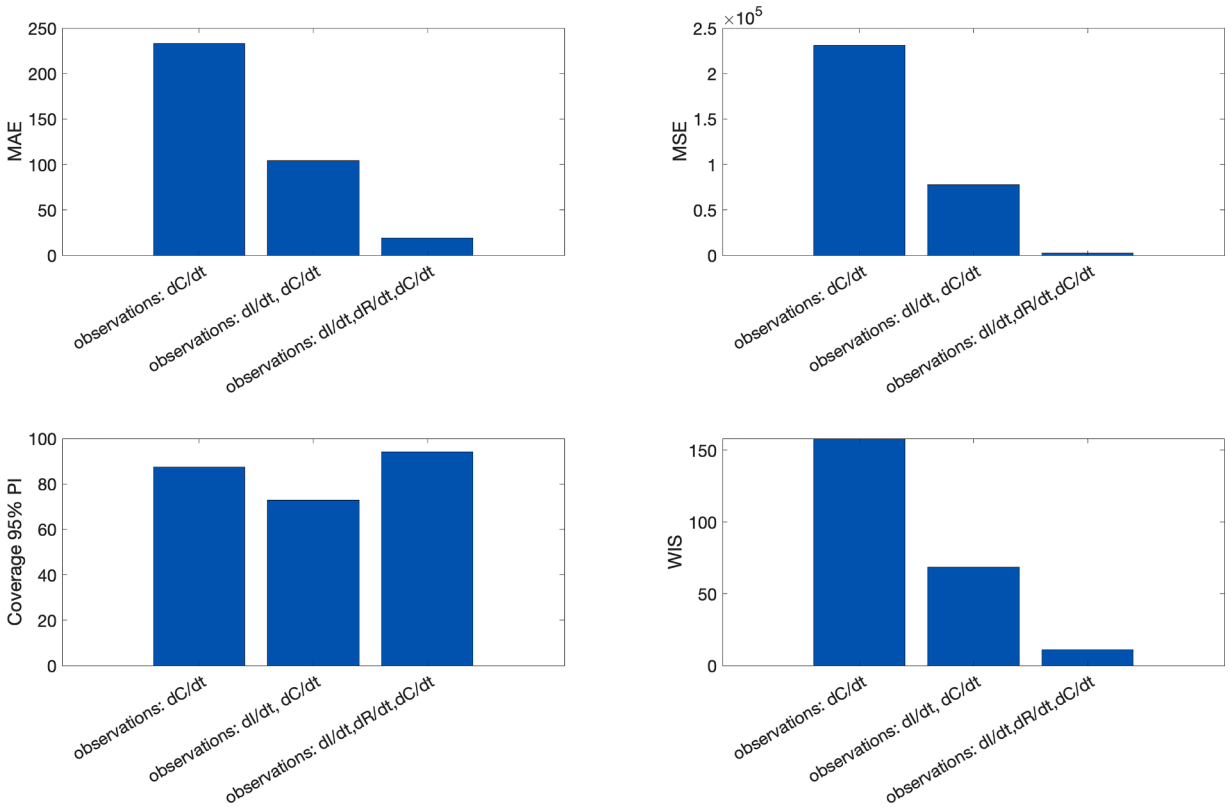


Fig. 7. Average performance metrics for 30-day ahead forecasts across different calibration periods as a function of the number of observed states used to calibrate the SEIR model. The figure demonstrates the clear relationship between the number of observed states and the accuracy of the model's forecasts, with more observed states consistently leading to improved performance metrics, such as reduced error and narrower prediction intervals. This underscores the critical role of incorporating multiple data streams in model calibration to enhance both the precision and reliability of epidemic forecasts.

mathematical models. For instance, the generalized-growth model (GGM) allows researchers to determine whether the epidemic is growing exponentially, as expected in a homogeneous mixing scenario consistent with the SEIR model and the underlying assumption supporting the definition of the basic reproduction number (R_0) [33,111]. Conversely, a slower, sub-exponential growth phase could indicate that additional factors, such as the mode of transmission, behavioral changes, or the impact of social distancing measures, influence the early dynamics and reduce transmission rates. In such a scenario, the effective reproduction number, R(t), declines over time [112].

Our analysis shows that the GGM fits Spain's first 25 days of the COVID-19 pandemic well, with a growth scaling parameter close to 1.0, indicating near-exponential growth dynamics (Fig. 8). This is consistent with early transmission stages of respiratory airborne pathogens like seasonal influenza, where the disease spreads rapidly through a fully susceptible population. Importantly, we find that the GGM parameters are practically identifiable from the early growth phase, enhancing confidence in the robustness of these estimates. Given these findings, the SEIR model provides a reasonable initial framework for mechanistic modeling.

4.2. *Formulating a mechanistic model: SEIR model incorporating reported and unreported cases*

Structural identifiability. The structural identifiability analysis of the model, excluding initial conditions, using DAISY reveals that while parameters k and γ are structurally identifiable, β and ρ are not structurally identifiable (see [59]). The correlation between β and ρ, contributes to this lack of identifiability. However, we can assume that the initial number of infectious individuals (e.g., $I(0)$) corresponds to the first data point in the daily curve, $E(0) = 0, R(0) = 0,$ and $S(0)$ can be approximated by Spain's population size in the context of a novel pathogen. Knowledge of the initial conditions resolves the identifiability issue, ensuring that all parameters in the model become structurally identifiable. The next step is to evaluate the practical identifiability of the model parameters when the model is confronted with real data.

Practical identifiability. When estimated along the transmission rate (β) and keeping the latency and recovery parameters fixed, our results indicate that parameter ρ is not practically identifiable during the initial phase of the epidemic, as evidenced by the wide uncertainty associated with this parameter (Fig. 9). In fact, its 95 % confidence interval spans 66 % of its possible value range. This suggests that the available data from the early outbreak period are insufficient to reliably estimate ρ, likely due to limited observations

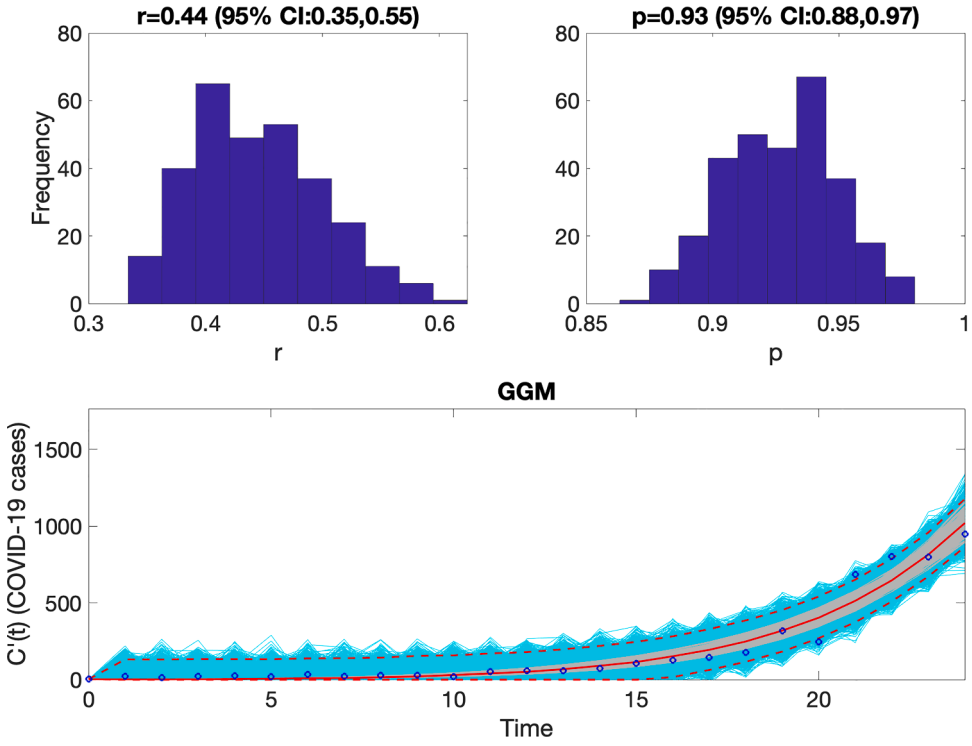


Fig. 8. Fitting the generalized-growth model (GGM) to the early phase of the COVID-19 epidemic in Spain with a normal error structure. The top panels display the empirical distribution of the estimated parameters alongside the model fit (solid line) to the observed case counts during the first 25 days of the epidemic (dots). The estimated growth scaling parameter ($p \approx 1.0$) indicates near-exponential growth, reflecting rapid transmission in the initial phase, which is consistent with the early spread of respiratory pathogens in a fully susceptible population. The model fit is shown in the bottom panel. The dashed lines indicate the 95 % prediction interval around the model predictions, the shaded area indicates the mean bootstrapped curves derived from the parametric bootstrapping process with 300 bootstrapped realizations, and the cyan lines indicate the predictive uncertainty around the model fit.

or correlations with other parameters in the model. As a result, the model requires revision to address this issue**.** One potential adjustment is to fix ρ at a plausible value based on prior knowledge or expert opinion. Given that initial estimates suggest $\rho \approx 1.0$, an alternative approach would be to revert to the simpler SEIR model, excluding the ρ parameter, and evaluate whether this simpler model structure adequately captures the epidemic dynamics.

Retaining the unidentifiable parameter ρ in the model may introduce large uncertainties, reducing the precision of parameter estimates and the accuracy of forecasts. This can lead to misleading conclusions and less reliable guidance for public health interventions. Nevertheless, as shown below, the identifiability of ρ improves in the later stages of the epidemic as more data accumulates, particularly as the epidemic approaches its peak. During this period, changes in transmission dynamics may provide sufficient data to accurately estimate ρ and other parameters related to the effects of interventions, thereby improving model reliability and reducing forecast uncertainty.

Next, we consider a simpler SEIR model in our analysis of the initial growth phase of the epidemic to resolve the practical identifiability issue.

4.3. Revising the model: consider the simple SEIR model

As previously demonstrated, the simple SEIR model ensures that all epidemiological parameters are structurally identifiable (see [59]). By fixing the latency and infectious periods, we estimated the transmission rate during the initial phase of the COVID-19 epidemic in Spain. The model produced a good fit to the observed data, and the transmission rate was found to be practically identifiable (Fig. 10). The corresponding empirical distribution and estimate of the basic reproduction number is shown in Fig. 11. Importantly, the uncertainty associated with the transmission rate estimates was significantly reduced in this simpler model compared to the model that incorporates a reporting proportion while still maintaining a high quality of fit to the data. This highlights the advantage of using a less complex model for early-phase epidemic analysis, as it ensures a more robust and practically identifiable model.

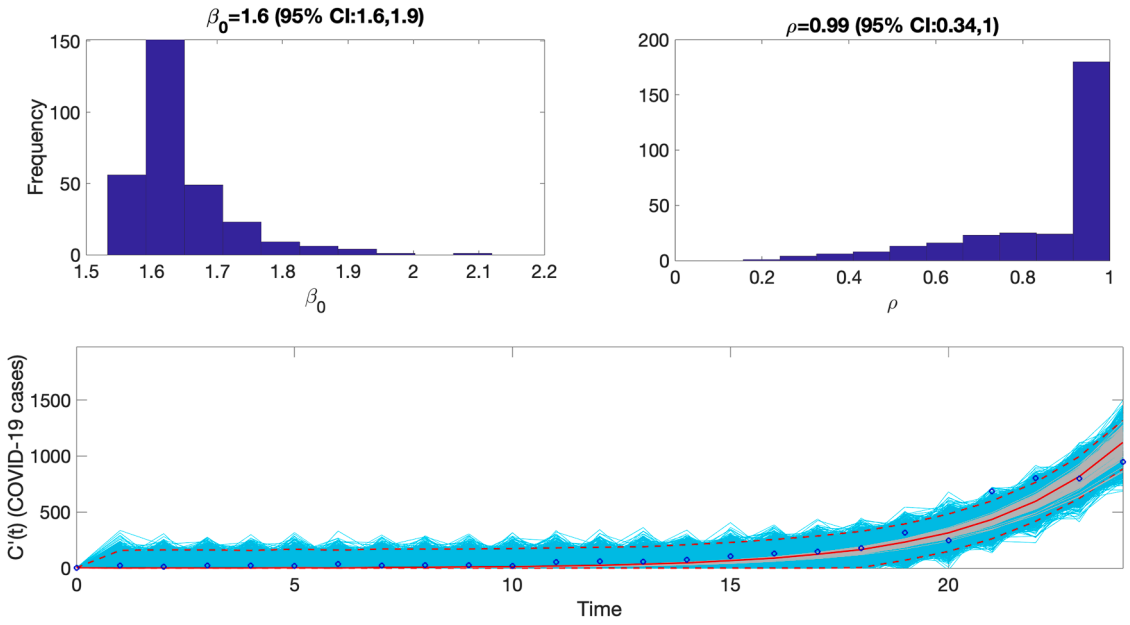


Fig. 9. Fitting the SEIR model with underreporting to the early phase of the COVID-19 epidemic in Spain using a normal error structure. The top panels display the empirical distribution of the estimated parameters and the model fit (solid line) to the observed case counts over the first 25 days of the epidemic (dots). Our results indicate that parameter ρ is not practically identifiable during the initial phase of the epidemic, as evidenced by the wide uncertainty associated with this parameter. The model fit is shown in the bottom panel. The dashed lines indicate the 95 % prediction interval around the model predictions, the shaded area indicates the mean bootstrapped curves derived from the parametric bootstrapping process with 300 bootstrapped realizations, and the cyan lines indicate the predictive uncertainty around the model fit.

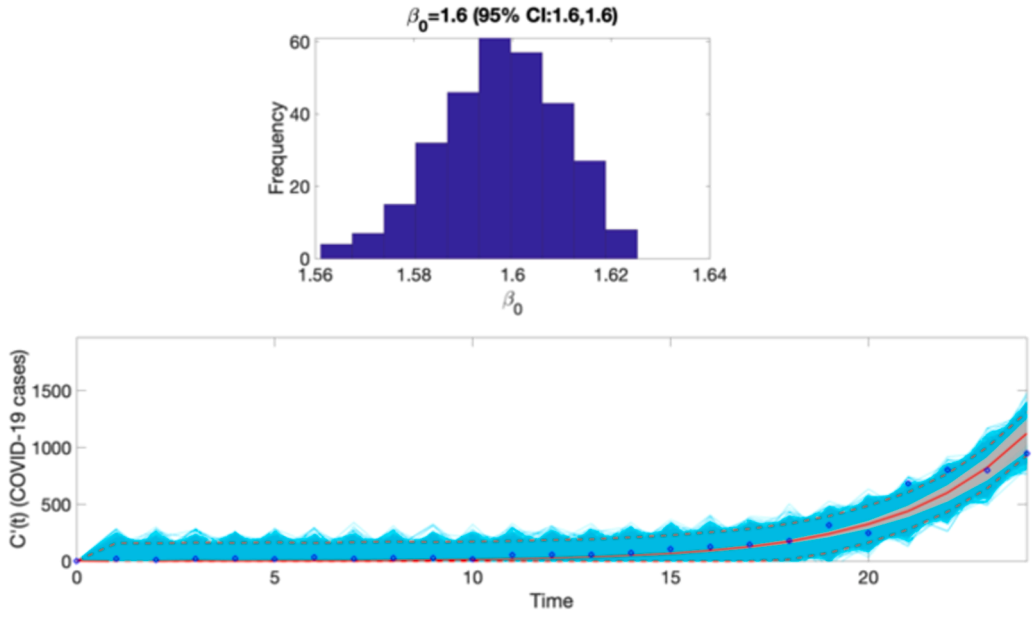


Fig. 10. Fitting the simple SEIR model to the early phase of the COVID-19 epidemic in Spain using a normal error structure. The top panel shows the empirical distribution of the transmission rate. Importantly, the uncertainty in the transmission rate estimates is significantly reduced in this simpler model, compared to the model with underreporting (Fig. 9), while still maintaining the quality of fit to the observed data. The bottom panel displays the model fit to the observed case counts over the first 25 days of the epidemic (dots). The dashed lines indicate the 95 % prediction interval around the model predictions, the shaded area indicates the mean bootstrapped curves derived from the parametric bootstrapping process with 300 bootstrapped realizations, and the cyan lines indicate the predictive uncertainty around the model fit.

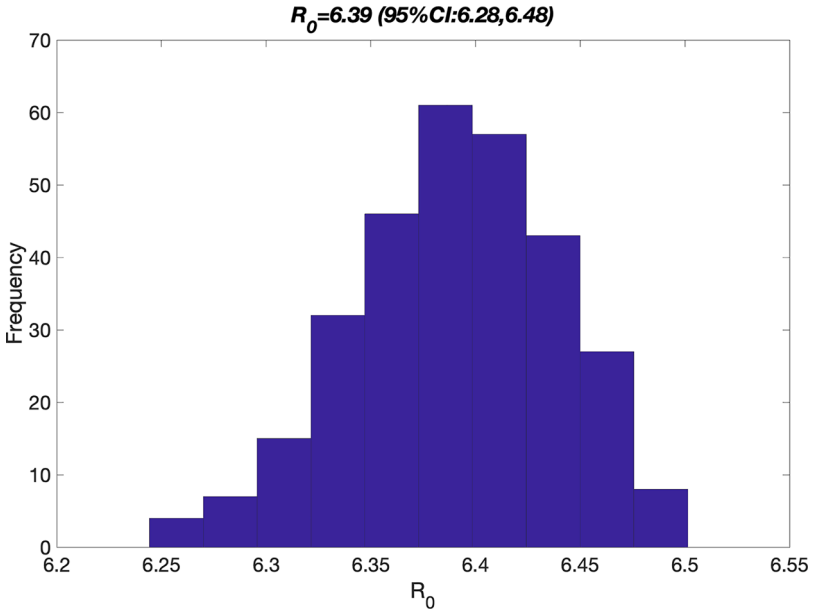


Fig. 11. The empirical distribution of the basic reproduction number (R_0) derived by fitting the SEIR model to the early growth phase of the COVID-19 pandemic in Spain.

4.4. Assessing the observation error

Since the model demonstrated a good fit during the early epidemic phase, we further explored the impact of different error structures on model performance. Specifically, we evaluated model fit using three error structures: Poisson, negative binomial, and normal. To compare these error structures, we analyzed four key metrics: MAE, MSE, coverage of the 95 % prediction interval, and the WIS (weighted interval score), as summarized in Table 1.

Our analysis demonstrates that the normal error structure performs well across all performance metrics (Table 1), providing superior coverage of the 95 % prediction interval and yielding the lowest MAE and WIS values. In contrast, the Poisson and negative binomial error structures exhibit higher error rates and lower coverage—particularly the Poisson structure, which achieves only 16 % coverage of the 95 % prediction interval. Overall, the normal error structure is a robust choice for this model, effectively balancing precision and accuracy in fitting the early epidemic data.

Since the early growth phase of the epidemic exhibits near exponential behavior, we can explore using the SEIR model enhanced with a power-law scaling exponent (α) to more accurately capture the non-homogenous transmission dynamics [32]. Indeed, Fig. 12 shows that the additional α parameter is well identified along the transmission rate from the early growth phase. In the next section, we consider an extended SEIR model that incorporates the effects of social distancing interventions during the COVID-19 pandemic in Spain.

4.5. SEIR model with interventions for the COVID-19 pandemic in Spain

We can employ a more elaborate SEIR model incorporating a dynamic transmission rate to capture the effect of interventions during the later stages of the epidemic in Spain. The model assumes that interventions, such as social distancing or lockdowns, gradually reduce the transmission rate from an initial value β_0 to a lower value β_1. This decline is modeled using an exponential decay function over time, reflecting how the transmission rate decreases as interventions are implemented:

$$\beta(\mathrm{t}) = \beta_1 + (\beta_0 - \beta_1) \cdot \mathrm{e}^{-\ qt}$$

where:

Table 1
Calibration performance of the SEIR model fitted to the early growth phase of the COVID-19 pandemic in Spain using *QuantDiffForecast*. A normal error structure yields the best performance metrics.

	Calibration period	MAE	MSE	Coverage 95 % PI	WIS
Normal	25	46.81	5983.32	92.00	31.44
Poisson	25	50.31	7237.97	16.00	43.34
NB	25	49.94	7020.39	76.00	32.19

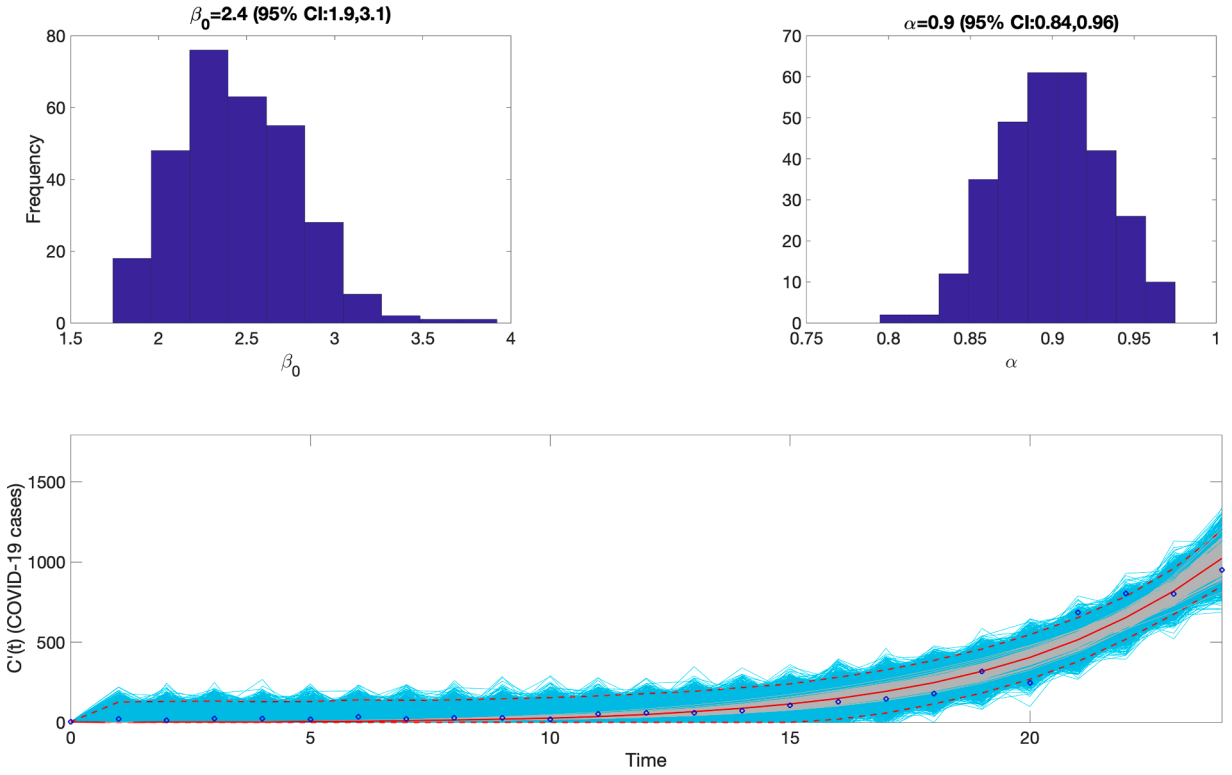


Fig. 12. Practical identifiability analysis of the SEIR model with non-homogenous mixing parameter α **.** The top panels display the empirical distribution of the estimated parameters and the model fit (solid line) to the observed case counts over the first 25 days of the epidemic (dots). Our results indicate that parameter α is well-identified during the initial phase of the epidemic. The model fit is shown in the bottom panel. The dashed lines indicate the 95 % prediction interval around the model predictions, the shaded area indicates the mean bootstrapped curves derived from the parametric bootstrapping process with 300 bootstrapped realizations, and the cyan lines indicate the predictive uncertainty around the model fit.

- β_0 the initial transmission rate before interventions.
- β_1 is the reduced transmission rate after interventions.
- q controls the speed of the decline in the transmission rate.

This model is particularly suitable for application as interventions take hold and more data accumulates during the later stages of the epidemic. As the epidemic progresses, the effects of interventions and changes in transmission dynamics become clearer, allowing for better parameter estimation and a more accurate representation of the disease's trajectory. By incorporating both the impact of

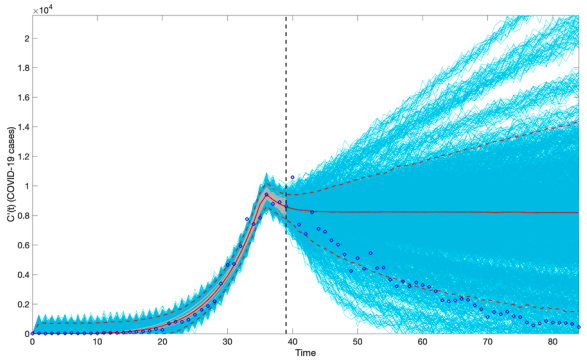


Fig. 13. . A 45-day forecast was generated after calibrating the model, which incorporated interventions and non-homogeneous mixing, using data from the first 40 days of the COVID-19 pandemic in Spain. The plot displays the model fit (solid line) to the observed case counts (dots) over the first 100 days of the epidemic. The dashed lines indicate the 95 % prediction interval around the model predictions, the shaded area indicates the mean bootstrapped curves derived from the parametric bootstrapping process with 300 bootstrapped realizations, and the cyan lines indicate the predictive uncertainty around the model fit.

interventions and non-homogeneous mixing, this model provides a flexible framework for predicting the epidemic's progression as more information becomes available.

We explored using the SEIR model with interventions and non-homogeneous mixing, to derive 45-day-ahead forecasts using progressively longer calibration periods. The calibration and forecasting performance of the model during the first COVID-19 wave in Spain are showcased in Figs. 13–17. Fig. 13 illustrates the model's fit to observed case counts over the first 45 days, up until the peak of the first pandemic wave, and a 45-day ahead forecast with the 95 % prediction intervals and bootstrapped uncertainty estimates highlighting the model's predictive accuracy. The corresponding empirical distributions and parameter estimates, which are well identified from the data, are shown in Fig. 14.

Using data from the entire first wave (100 days) of the COVID-19 pandemic in Spain, Fig. 15 illustrates the model's fit to observed case counts, with the 95 % prediction intervals and bootstrapped uncertainty estimates highlighting the model's predictive accuracy. Parameter estimates, shown in Fig. 16, are well-identified, capturing key dynamics such as the transmission rate, intervention effects via the q parameter, and mixing heterogeneity through the parameter α. The 45-day-ahead sequential forecasts in Fig. 17 demonstrate improved accuracy with longer calibration periods, reflected by tighter prediction intervals and improved performance metrics (Table 2), particularly for forecasts made after the first 45 days of the epidemic curve (Fig. 17).

5. Advancing epidemic forecasting: work in progress

In recent years, machine learning (ML) and artificial intelligence (AI) have emerged as crucial tools for predicting natural events such as weather conditions and epidemic spread. These technologies enable researchers to analyze large datasets and produce more precise short-term predictions [113–121]. These methods can account for the nonlinear complexities of epidemic spread, which are often not explicitly observed in surveillance data [122–124]. For example, deep learning (DL) models have been highly effective in forecasting short-term trends in COVID-19 cases, hospitalizations, and deaths, proving crucial in public health planning and response efforts [118].

Despite these advances, ML models, especially deep learning architectures, often fail to provide the same level of biological interpretability as mechanistic models [125–127]. While deep learning has advanced short-term forecasting, its black-box nature poses a significant limitation—it does not inherently reveal the underlying epidemiological transmission dynamics, which are essential for guiding public health interventions. As a result, there is increasing recognition of the need to integrate ML-based models with dynamical systems approaches to improve both predictive accuracy and interpretability [128]. Merging the data-driven strengths of AI with the theoretical foundations of mechanistic models could enable the development of hybrid models that offer more accurate epidemic forecasting and provide deeper insights into disease transmission mechanisms.

The workflow outlined in this paper highlights the critical role of identifiability—both structural and practical—in ensuring precise parameter estimation and reliable forecasts. Introducing ML into this workflow adds a layer of complexity, particularly in maintaining the identifiability and interpretability of hybrid models that combine data-driven and mechanistic approaches [129–133]. Since purely data-driven ML models often obscure mechanistic insights into system dynamics, future research should focus on developing methods to more effectively integrate these techniques. Thus, a key challenge lies in ensuring parameter identifiability within these hybrid models—a non-trivial task that requires further investigation. Potential strategies may involve applying regularization techniques or embedding mechanistic constraints within ML architectures, to help preserve identifiability and enhance both predictive accuracy and

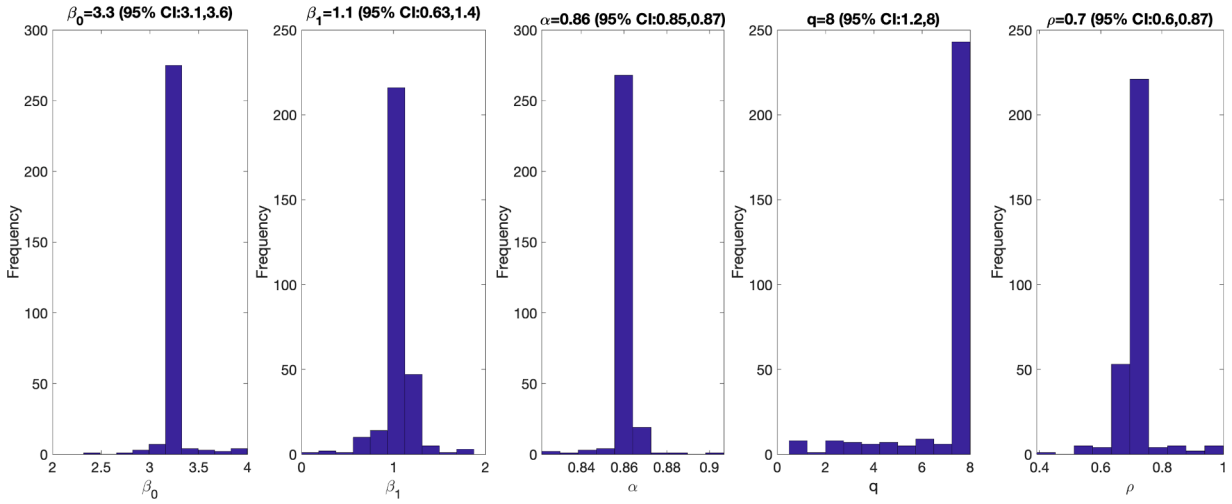


Fig. 14. Parameter estimates and their empirical distributions derived from the model incorporating interventions and non-homogeneous mixing for the COVID-19 epidemic in Spain. The estimates include key epidemiological parameters such as the transmission rate, intervention effect, and mixing heterogeneity using the first 40 days of the COVID-19 pandemic in Spain. Overall, the parameters from the data are practically well-identified.

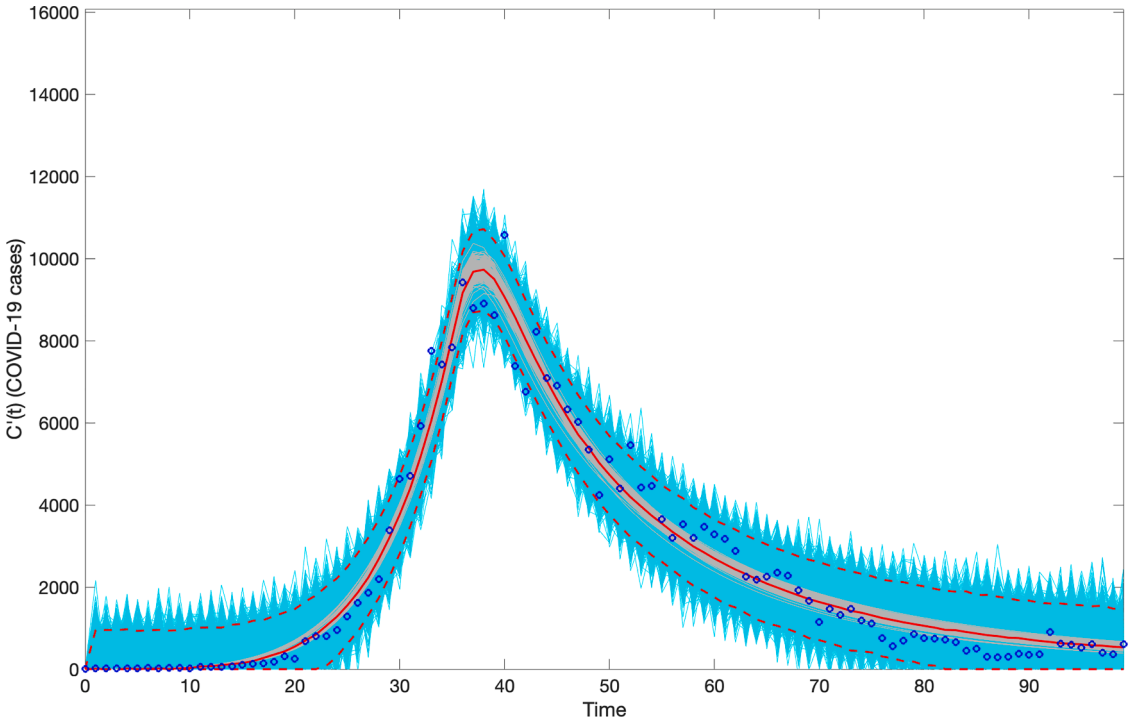


Fig. 15. Fitting the model incorporating interventions and non-homogeneous mixing to the first 100 days of the COVID-19 pandemic in Spain. The plot shows the model fit (solid line) to the observed case counts (dots) over the first 100 days of the epidemic. The dashed lines indicate the 95 % prediction interval around the model predictions, the shaded area indicates the mean bootstrapped curves derived from the parametric bootstrapping process with 300 bootstrapped realizations, and the cyan lines indicate the predictive uncertainty around the model fit.

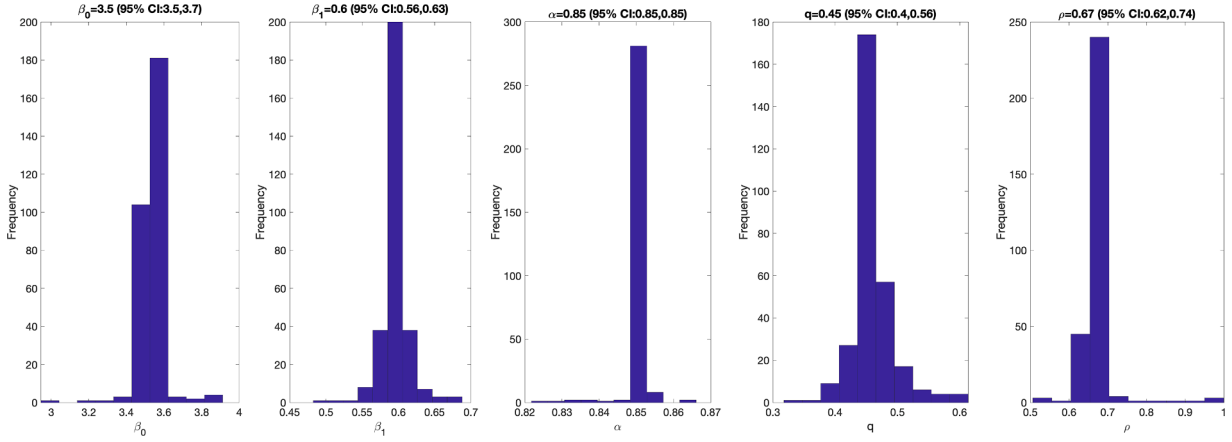


Fig. 16. Parameter estimates and their empirical distributions derived from the model incorporating interventions and non-homogeneous mixing for the COVID-19 epidemic in Spain. The estimates include key epidemiological parameters such as the transmission rate, intervention effect, and mixing heterogeneity over the 100 days. Overall, the parameters from the data are well-identified.

biological interpretability.

By combining the flexibility of ML with the interpretability and theoretical rigor of dynamical systems, there is substantial potential to improve forecast accuracy and offer more actionable insights for public health decision-making. However, significant challenges remain in ensuring hybrid models adhere to rigorous workflows for parameter identifiability and calibration—key elements for reliable epidemic forecasting. Addressing these challenges will require innovative approaches that integrate the data-driven capabilities of ML with the structural clarity of mechanistic models. This integration must ensure that the resulting framework not only improves predictive performance but also maintains the transparency and interpretability essential for public health applications.

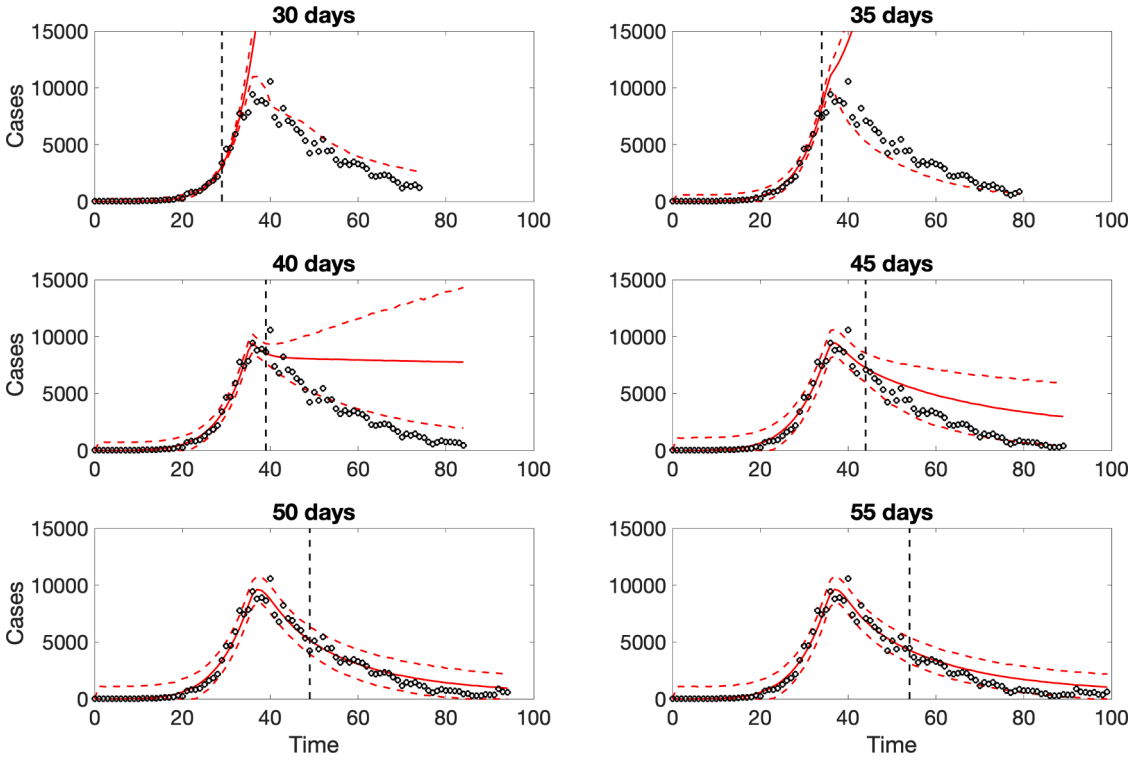


Fig. 17. 45-day ahead forecasts using an increasingly longer calibration period for fitting the SEIR model that incorporates interventions and non-homogeneous mixing during the first COVID-19 wave in Spain. The solid line represents the model fit and forecast. The dashed lines denote the 95 % prediction interval, capturing the uncertainty of future projections. The vertical dashed line separates the calibration from the forecasting period.

Table 2
Forecasting performance metrics for the 45-day ahead forecasts derived from the SEIR model incorporating interventions and non-homogeneous mixing, evaluated across progressively longer calibration periods during the first COVID-19 wave in Spain.

Calibration period	MAE	MSE	Coverage 95 % PI	WIS
30	338,882.49	256,601,419,414.54	11.11	219,960.64
35	46,790.81	3,460,948,859.79	86.67	30,063.53
40	4834.83	27,638,183.18	17.78	3535.65
45	1484.19	2,545,303.36	100.00	879.22
50	350.22	187,782.56	97.78	232.24
55	669.70	550,589.78	97.78	414.40

6. Methodology of molecular surveillance of epidemiological dynamics

6.1. *Major challenges and opportunities in molecular surveillance of viral diseases*

Tracking and forecasting infectious disease dynamics is a crucial public health task. While traditional surveillance methods, such as surveys, contact tracing, and data collection on exposures and risk factors, are widely used, they are often labor-intensive, time-consuming, and prone to biases. Moreover, the accuracy of these methods can be compromised by sampling and reporting biases.

Recent breakthroughs in genomics and high-throughput sequencing technologies have spurred the rapid development of *genomic epidemiology* and *genomic surveillance* (see Fig. 18). These multidisciplinary fields leverage viral genome data to infer and forecast epidemiological dynamics [134,135]. As viruses evolve quickly through mutations and genomic alterations, these changes — when accurately sampled and analyzed — can offer valuable insights into the history of viral spread.

Modern high-throughput sequencing technologies can rapidly generate vast amounts of viral genomic data for surveillance and outbreak investigations. With decreasing sequencing cost and the growing capabilities of computational technologies, genomic epidemiology has become more feasible and affordable. This presents a unique opportunity to observe viral evolution in near real time, significantly enhancing genomics-based forecasting methods.

The power of genomic methods was vividly demonstrated during the COVID-19 pandemic caused by SARS-CoV-2. Shortly after the initial outbreak in Wuhan, China, researchers sequenced and assembled the virus's genome, identifying it as a novel betacoronavirus

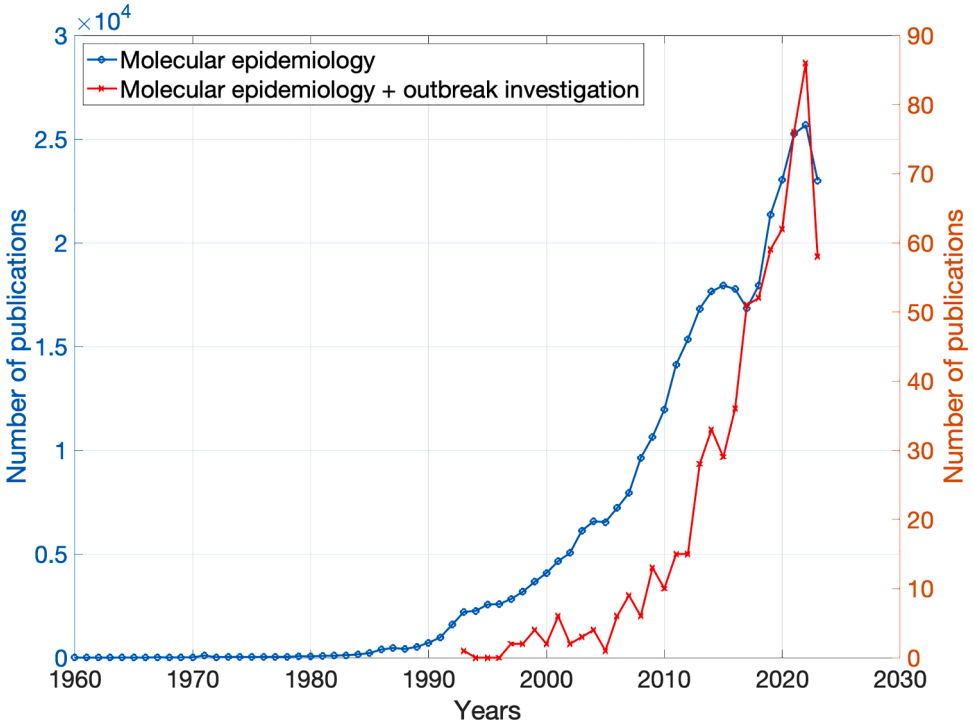


Fig. 18. Trends in publications related to molecular epidemiology and those focused on outbreak investigations. The trends indicate that the application of molecular epidemiology in outbreak investigations started to take off around 2005, coinciding with the regular use of sequencing technologies in biology and epidemiology. Data for this figure were retrieved from the Web of Science.

closely related to SARS-CoV-1 and bat coronaviruses, suggesting a zoonotic origin. By the time the World Health Organization (WHO) declared a Public Health Emergency of International Concern, over 300 SARS-CoV-2 genomes had already been sequenced. At the pandemic's peak, nearly 70,000 genomes were being uploaded daily to public repositories such as GISAID, COG-UK, and GenBank [136]. This unprecedented volume of genomic data enabled rapid tracking of transmission routes, early detection of Variants of Concern (VOCs) and Variants of Interest (VOIs), and, for the first time, reliable forecasting of viral evolutionary dynamics and their epidemiological impacts.

A major advantage of genomic surveillance is its ability to avoid many of the biases inherent in traditional epidemiological methods. However, it also presents unique challenges, often stemming from the complexity of genomic data and the limitations of sequencing technologies. Viral genomes do not directly encode epidemiological history; thus, advanced computational methods are required to infer it. The accuracy of these inferences depends heavily on the quality of the underlying mathematical and statistical models, making model selection a critical factor in the success of genomic epidemiology methods applied for each particular scenario. More complex models can capture intricate epidemiological dynamics and offer deeper, more accurate insights. However, these models often demand significant computational resources and may struggle to scale with even moderately sized datasets. Consequently, the effective application of genomic epidemiology requires striking a delicate balance between biological realism and computational efficiency.

Moreover, sequencing technologies are inherently imperfect, introducing errors and noise into the data that can affect downstream analyses. The highest sequencing coverage is typically achieved through short-read technologies, which generate fragmented sequences (reads) rather than full-length genomes. Before any epidemiological inferences can be drawn, these sequencing errors must be corrected, and the full genomes assembled using specialized bioinformatics tools—a process commonly known as primary processing. Primary processing of viral genomic data is particularly challenging due to the high heterogeneity of viral populations, where signals from minor variants can be difficult to distinguish from sequencing noise. The computational toolkit for viral genomics is continually evolving, with frequent updates designed to address emerging challenges. As a result, selecting the appropriate bioinformatics tools for primary analysis is crucial.

Until recently, the relatively high cost and logistical challenges of sample collection were significant obstacles to the efficient use of genomic surveillance. However, recent advances in wastewater-based surveillance offer a more cost-effective and efficient sampling method. The detection of trace viral genomic material in wastewater has been successfully used to track SARS-CoV-2 infection dynamics and has been implemented in many countries worldwide [136]. Wastewater-based surveillance accounts for both symptomatic and asymptomatic cases, can detect viral variants before they become widespread, and provides more balanced estimates of viral prevalence compared to clinical testing alone [136].

Finally, it is important to recognize that traditional epidemiological and genomic approaches are not mutually exclusive; they are

most effective when used together. The most accurate genomic epidemiology approaches often integrate case-specific epidemiological evidence, which can serve as constraints on model parameters or inform priors in Bayesian frameworks. This integration helps resolve inconsistencies and refine evolutionary scenarios that may arise when relying solely on genomic data.

6.2. Genomic methods and models for detection of transmission clusters and outbreaks

A key application of genomic epidemiology is the investigation of viral outbreaks [134,135]. This process typically involves two major steps: (a) identifying transmission clusters and (b) inferring transmission histories within these clusters. Numerous bioinformatics tools have been developed for these tasks, including, but not limited to, Outbreaker and Outbreaker 2 [137,138], SeqTrack [139], SCOTTI [140], SOPHIE [141], Phybreak [142], Bitrugs [143], BadTrIP [144], Phyloscanner [145], StrainHub [146], TransPhylo [147,148] and its extension TransPhyloMulti [149], STraTUS [150], TreeFix-TP [151], QUENTIN [152], VOICE [153], HIVTrace [154], GHOST [155], MicrobeTrace [156], SharpTNI [157], TiTUS [158], TNeT [159], AutoNet [160], SMiTH [161], and others [148,162–169]. These tools have been successfully employed to investigate outbreaks and track transmission dynamics of a variety of pathogens [170–175].

From a methodological standpoint, transmission inference methods can be broadly classified into network-based and phylogenetic approaches. Network-based methods have gained significant traction, particularly among researchers studying HIV and HCV, and have been adopted as a standard approach for outbreak investigations by the CDC [154,156,169,173,175–177].

This approach typically involves two stages. First, a relatedness graph is constructed, where vertices represent infected individuals and edges connect those whose intra-host viral populations are closely related, based on a chosen population genetics measure. In some cases, this graph provides sufficient information for epidemiological analysis [169,176,177]. However, it usually contains edges that do not correspond to direct transmissions. Therefore, in an optional second stage, a transmission tree is inferred, usually as a spanning tree of the genetic relatedness graph, optimized according to a selected objective. While network-based methods cannot determine the directionality of transmissions [178], they are highly scalable, simple to implement, and produce results that are both visualizable and interpretable. Several tools, such as GHOST [155] and MicrobeTrace [156], offer web-based platforms with user-friendly interfaces.

Another family of methods utilizes a phylogenetic approach, where transmission networks are reconstructed from viral phylogenetic trees. In some cases, the topology of the phylogenetic tree alone can reveal transmission events without requiring further calculations. For example, a paraphyletic relationship between intra-host viral populations—where populations intermingle rather than form distinct clades—indicates recent transmission between the corresponding hosts [153,179–181].

However, in general a phylogenetic tree does not directly indicate transmission events. Internal nodes in the tree represent lineage divergences [13], some of which arise from transmissions between hosts, while others occur within already infected hosts. Therefore, deriving a transmission network from a phylogenetic tree requires inferring the traits of internal nodes by annotating them with labels representing the infected hosts. These labels allow researchers to distinguish whether each divergence event occurred within a host or resulted from the transmission of viral variants to a new host. Trait inference is typically framed as a character optimization problem, where traits (or characters) are reconstructed to maximize or minimize a predefined objective function.

Various objectives and constraints have been implemented in existing methods. The simplest approaches are based on the principle of maximum parsimony, which favors transmission histories with fewer transmission events, smaller bottleneck sizes, or minimal back-transmissions as the most plausible solutions [139,145,150,158,159,182]. These methods are relatively scalable, and the associated algorithmic problems can often be efficiently solved using dynamic programming or, in more complex scenarios, through Integer Linear Programming (ILP) [183], with ILP solvers offering reasonable computational efficiency.

Maximum likelihood and Bayesian phylogenetic models [137,140,142,144,149], which incorporate constraints regularized as priors, provide a more nuanced biological and epidemiological perspective. These models allow for the inclusion of temporal information to improve the accuracy of transmission link reconstructions [184], while also incorporating parameters that describe evolutionary and epidemiological dynamics, such as effective reproduction numbers and sampling rates. However, a drawback of these parameter-rich models is their computational complexity, which often leads to challenging optimization problems.

It is important to note that when relying solely on genomic data, multiple optimal solutions may fit the observed data equally or nearly as well. To resolve these ambiguities, additional epidemiological evidence can be incorporated to establish the order of infections and rule out unlikely transmission links [137,140,142,143,147,158,163], provided such evidence is available. This evidence can include exposure intervals, where hosts with overlapping intervals are more likely to be connected, as well as symptom onset, diagnosis, or sample collection dates, which can help estimate infectious periods if their distribution is known or reasonably assumed. In rare cases, known contact networks can also be used to refine transmission inferences [175,185].

For pathogens associated with endemic diseases and chronic infections, such as HIV or Hepatitis C, case-specific epidemiological data is often unavailable, non-informative, or sensitive [141,152,169]. In such cases, general knowledge about the expected properties of viral transmission networks—derived from the structure of susceptible populations—can serve as a substitute [141,152]. Infectious diseases typically spread through social networks of contacts, and transmission networks often mirror the underlying properties of these social structures [177,186–188]. The general characteristics of these social networks (e.g., being scale-free) are well established in network theory, sociology, and classical epidemiology [189]. This allows the transmission inference problem to be framed as finding transmission networks that are consistent with both the observed genomic data and the highest probability of being subnetworks of random contact networks [141,152].

An example of transmission network inference methodology in evaluating epidemiological dynamics is the study of a large 2015–2016 HIV/HCV outbreak in rural Indiana. Identified by the Indiana State Department of Health (ISDH) in early 2015, the outbreak in Scott County led to hundreds of HIV and HCV infections and a public health emergency [190]. It was traced to unsafe

injection of the opioid oxymorphone [191], highlighting the link between viral transmission and the opioid abuse epidemic [192,193].

In ref. [141], the transmission detection tool SOPHIE was combined with phylogenetic tree construction to infer and analyze the HCV transmission network in this outbreak. SOPHIE (SOcial and PHilogenetic Investigation of Epidemics) infers transmission networks from viral phylogenetic trees and models inter-host social networks as random graphs with expected degree distributions. It samples from the joint distribution of phylogeny ancestral traits, estimates the probability that the sampled networks are subgraphs of a random contact network, and summarizes them into a consensus network.

In the study, genomic data collected by the CDC during the outbreak investigation were used to construct a phylogenetic tree with RAxML [194]. The tree was then post-processed using TreeTime [195] to estimate the timing of branching events. This time-scaled tree was used as input for SOPHIE, which inferred transmission networks and their probabilities, providing estimates for the timing of transmissions. Importantly, SOPHIE generates a distribution of possible transmission networks, allowing confidence intervals to be calculated for transmission times and derived epidemiological parameters.

SOPHIE's output was used to estimate key epidemiological dynamics, including incidence, generation times, and the effective reproduction number. Incident case numbers, i.e. numbers of transmissions within specific time intervals, were calculated directly. The incidence curve suggests the outbreak began in mid-2012, entering an exponential growth phase in 2014 (Fig. 19(c)), before rapidly declining after the public health emergency was declared. The exponential phase largely coincides with the HIV spread in the same community [175]. Additionally, the fact that HIV co-infected hosts formed a connected subnetwork within the HCV transmission network suggests both outbreaks were driven by the same epidemiological mechanism, with HCV preceding HIV by several years, and HIV spread facilitated by the pre-existing HCV network.

Similar to incidence, generation times—i.e., the intervals between infection events—were derived from the timed phylogenies and internal node traits corresponding to transmission events (Fig. 19b). These, along with incidence data, were processed using EpiEstim [15] to estimate the effective reproduction number R_t over a 1-month sliding window during the exponential phase. The results show that before the public health emergency, R_t ranged from 1.81 to 2.33, indicating sustained transmission, but dropped below the epidemic threshold of 1 following the declaration (Fig. 19d).

These results also highlight how epidemiological parameter estimation based on inferred transmission networks offers significant advantages as they provide more realistic estimates than more traditional methods based on random mixing models [196]. In the study discussed, this approach produced more moderate and seemingly realistic reproduction number estimates compared to the higher values from the birth-death skyline phylodynamics model ($R_0 = 6.6$, 95 % CI: 3.2–9.9; $R_0 = 5.1$, 95 % CI: 1.7–9.2) as inferred by BEAST [175,197,198] (see the next section for the detailed discussion of phylodynamics methods). These estimates also align better with the $R_0 = 3.8$ obtained for the parallel HIV outbreak via contact tracing [175].

6.3. *Phylodynamic surveillance and epidemiological forecasting*

Bayesian phylogenetics and phylodynamics are key tools in genomic epidemiology. Phylodynamics is a statistical approach that estimates pathogen evolutionary and population dynamics from genomic data [199] by fitting a predefined epidemiological and evolutionary models to viral phylogenetic trees. Fitting is carried out via sampling from the posterior distribution of model parameters. These methods are invaluable when traditional epidemiological data (e.g., incidence statistics) are unavailable or unreliable, or when genomic data are insufficient to reconstruct actual transmission links, as described earlier.

Phylodynamic models are typically framed within a Bayesian framework and can be formulated as follows. It is assumed that observed viral genomes $G = \{g_1,\ldots,g_n\}$, their frequencies $A = \{a_1,\ldots,a_n\}$ and their time-labelled phylogenetic tree T result from the combined effects of mutation, selection, and transmission processes. This process is described by a model that is parametrized by epidemiological parameters E (e.g. effective reproduction number, becoming noninfectious rate, sampling rate, etc.), selection parameters F (e.g. fitnesses of viral variants reflecting their relative infectivity and/or transmissibility) and *mutation parameters* Θ (e.g. molecular clock rate and substitution rates). The model often utilizes Markov [200] or Poisson [168,201] models for genome evolution, alongside birth-death [202,203], coalescent [204], or replicator models [205] for phylogeny construction. The model is often described by a system of coupled ODEs, and the goal is to estimate parameters and the phylogenetic tree by sampling from or maximizing the posterior probability distribution:

$$p(T,F,\Theta,E|G,A) \tag{1}$$

which is decomposed in a Bayesian manner [76,66,80,37] as

$$p(T,F,\Theta,E|G,A) \propto p(G,A|T,E,F,\Theta)\cdot p(T|F,E)\cdot p(E)\cdot p(F)\cdot p(\Theta) \tag{2}$$

Inference of both discrete and continuous parameters is typically performed using Markov Chain Monte Carlo (MCMC) sampling from the distribution (**2**). MCMC explores the parameter space by generating Markov chains whose states represent possible parameter values, with transitions based on their relative likelihoods. By constructing a large sample of parameter sets, MCMC provides a detailed picture of the distributions of evolutionary and epidemiological parameters, including medians and confidence intervals, while accounting for uncertainties and noise. For inference, the parameter sets must be finite. To capture changes over time, parameters can be approximated as piecewise continuous functions over a predefined number of disjoint time intervals (so-called *skyline models*). The interval lengths can be uniform or set according to specific epidemiological conditions.

From this description, it is clear that the successful application of phylodynamic methods depends heavily on the careful selection of evolutionary and epidemiological models, parameter priors, and MCMC transition probabilities. This task is typically left to the user,

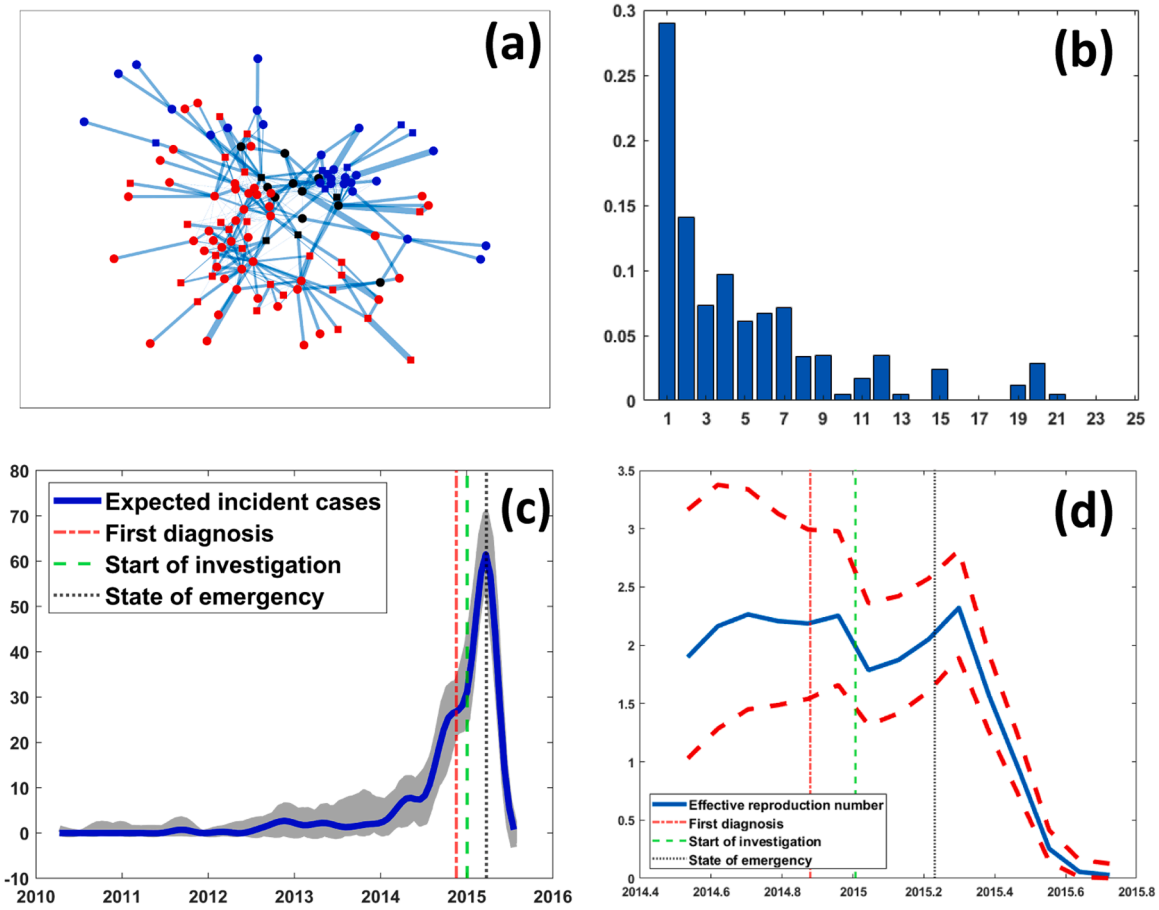


Fig. 19. Computational analysis of the Indiana HCV outbreak.
(a) Consensus transmission network, with edge thickness proportional to inferred likelihood support (only edges with support above 0.0005 are shown). Red, blue and black dots are patients infected with HCV subtypes 1,3 and both. (b) Distribution of generation times (in months). (c) Dynamics of incident cases over time, with vertical lines marking key public health events. (d) Effective reproduction number during the exponential stage, with vertical lines indicating major public health interventions.

but phylodynamic tools often provide a selection of popular options. It is important to note that phylodynamic inference is computationally intensive—especially for certain models, where even likelihood calculations [202,206], let alone MCMC sampling, can be demanding. Adequate computational resources are essential when analyzing large numbers of pathogen sequences.

The most widely used phylodynamics tool among researchers and public health practitioners is BEAST (Bayesian Evolutionary Analysis Sampling Trees) [198], which features a graphical interface, a wide range of models and parameters, and libraries to speed up calculations. Its modular design also allows easy integration of new models. Other tools include phylodyn [207], RevBayes [208] (Bayesian methods), and TreeTime [195] and TreePar [197] (maximum likelihood methods).

Phylodynamic methods can reveal past epidemiological dynamics from viral genomic data. A notable example is the study of the HCV epidemic in Egypt [215], where Bayesian skyline and birth-death skyline models inferred that the rapid rise in HCV infections was triggered by antischistosomal injection campaigns in the mid-20th century. In further analyses, the inferred effective population size and reproduction number curves can be extrapolated for future forecasting [209].

An example of such analysis is the author's study of COVID-19 dynamics in Belarus [210] and Ukraine [211], two neighboring countries with differing approaches to containment measures. Belarus implemented limited measures, while Ukraine followed broader non-pharmaceutical interventions (NPIs) similar to other European countries. Due to limited epidemiological data and undersampling in both countries [212], genomic data was used to replace missing epidemiological statistics. First, the Coalescent Bayesian Skyline (CBS) model was employed to infer historical population size changes and times of lineage splits. Priors for evolutionary parameters were set by using HKY+Γ nucleotide substitution model and a strict molecular clock. HKY+Γ is a nucleotide substitution model that accounts for differences in transition and transversion rates and incorporates rate variation across genomic positions using a gamma distribution. The molecular clock rate was also assumed to follow a gamma (Γ) distribution with the mean and the standard deviation taken from published measurements for SARS-CoV-2 [213,214]. The number of time segments for the skyline model (four) was set based on the epidemiological considerations, as there were four growth and decline periods of the first and second COVID-19 epidemic

waves. MCMC sampling was run for 3×10^7 iterations, sampling every 3×10^3 iterations and the initial 10 % "burn-in" iterations, thus allowing for sufficient statistical power.

The maximum clade credibility tree inferred by the Coalescent Bayesian Skyline (CBS) model (Fig. 20) was used to identify the largest within-country transmission clusters, which were further analyzed using the more parameter-rich Birth-Death Skyline (BDSKY) model [215]. BDSKY infers effective reproduction numbers, recovery rates, and changes in the infected population over time, assuming a birth-death process where "births" represent infections and "deaths" represent recoveries. Given the presence of co-circulating lineages, we employed a linked model where lineages evolve independently but share substitution model parameters, molecular clock rates, and reproduction numbers from the same priors. The same substitution model, molecular clock settings, and MCMC sampling as in the CBS analysis were used.

Since the BDSKY model is parameter-rich, we incorporated informative priors. For the sampling rate, we used a *Beta* distribution prior with parameters set to reflect the low proportion of sequenced cases. The origin of each cluster was given a normally distributed prior based on CBS estimates, and for the rate of becoming non-infectious, we used published estimates [184,213]. After MCMC sampling, we used the tool EpiInf [216] to reconstruct cumulative case count trajectories from the sampled trees and BDSKY parameters.

The analysis enabled a comparison of the effects of NPIs during the first COVID-19 wave in Belarus and Ukraine. While a moderate but statistically significant decrease in the effective reproduction number R_t was observed in Belarus after NPIs were implemented, the reduction was smaller compared to countries with stricter measures. Additionally, the similar estimates of R_t for both countries suggest that widespread violations of lockdown and distancing measures in Ukraine, along with limited governmental control, diminished the effectiveness of NPIs there. This highlights the importance of regional demographic and social factors in COVID-19 epidemiology, alongside NPIs. The study also showed that the true number of infections in Belarus by the end of May 2020 was likely about four times higher than reported, facilitating more accurate forecasting of future case counts (Fig. 20).

6.4. *Genomic methods and models for evolutionary forecasting*

Understanding the predictability of evolution is a fundamental challenge in the life sciences, with significant practical implications for viruses and other pathogens. Even a modest ability to predict pathogen evolution can significantly improve our capacity to forecast and control the spread of infectious diseases [217–220].

Over the past decade, genomic epidemiology has played an increasingly signifcant role in tracking viral evolution, including monitoring the emergence of new strains and antigenic drift. This approach has proven effective in forecasting the epidemiological dynamics of viruses such as influenza, Ebola, and SARS-CoV2, as well as informing the biannual selection of influenza A and B vaccine seed strains [221]. Notable evolutionary surveillance systems used by public health professionals include the WHO's Global Influenza Surveillance and Response System (GISRS) and Nextstrain, an open-source genomic platform that has been providing tools for phylogenetic analysis and virus spread visualization since 2015.

Genomic methods for evolutionary prediction received significant advancement during the COVID-19 pandemic. Successive waves of COVID-19 have been driven by emerging variants of interest (VOIs) or variants of concern (VOCs), which are associated with altered phenotypic traits such as increased transmissibility [222,223], antibody resistance, and immune escape [224–226]. Each genomic variant is defined as a phylogenetic lineage, characterized by a distinct combination of single amino acid variants (SAVs) and/or indels

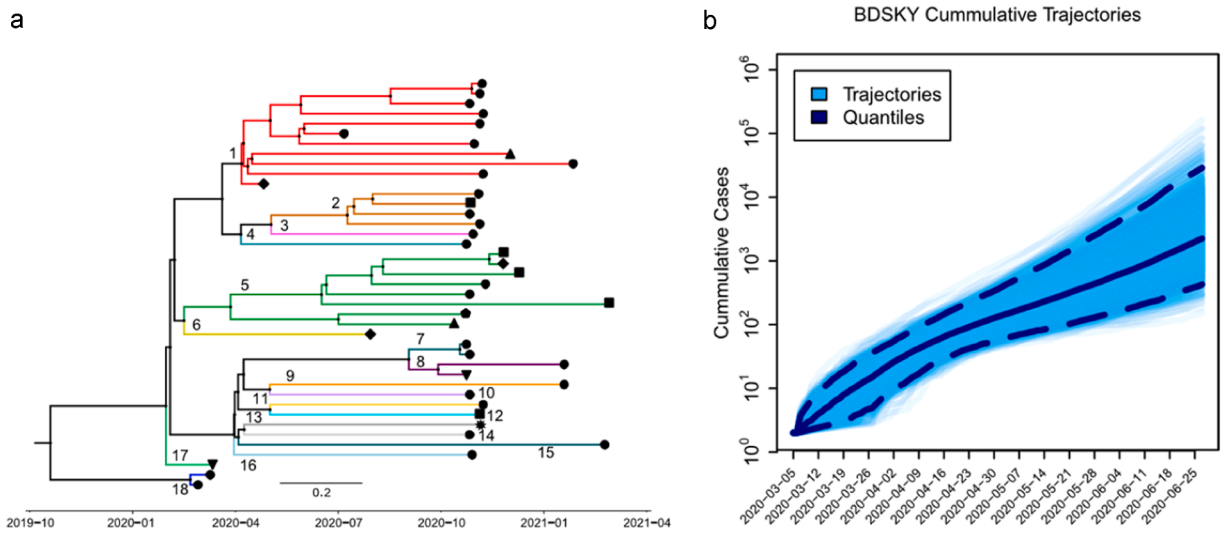


Fig. 20. Phylodynamics analysis of SARS-CoV-2 evolution and spread. Left: The annotated and timed phylogenetic tree for Belarusian SARS-CoV-2 sequences created using Coalescent Bayesian Skyline. Right: Possible cumulative case count trajectories. Solid blue and dashed lines represent a median and 95 % confidence intervals.

acquired during the evolution of SARS-CoV-2.

Given these developments, genomic surveillance methods have proven indispensable and effective for accurate COVID-19 epidemiological forecasting. Monitoring the evolution of SARS-CoV-2 has enabled the detection of spreading lineages and the prediction of their dynamics [136]. Tools like Nextstrain have been enhanced to estimate reproduction numbers and forecast viral lineage frequencies based on historical data, employing various regression techniques. However, despite their strengths, regression-based genomic surveillance methods have an important limitation — they are typically applied retrospectively. This means growing lineages and their fitness can only be assessed once they have already become sufficiently prevalent.

In contrast to regression-based forecasting, early detection methods aim to proactively identify SARS-CoV-2 variants that may become prevalent in the future, even if they have not yet emerged or are currently at low frequencies. This task is inherently more challenging [227], but several published methods address it using various machine learning techniques [217,218,228–232]. The largest group of these methods focuses on predicting the emergence of individual mutations, assuming either independent accumulation of mutations or that interaction effects can be averaged across different genomic backgrounds [218,231]. This can be framed as a classification problem, where mutations are categorized as "potentially spreading" or "non-spreading" based on carefully selected genomic, epidemiological, and physicochemical features.

Other methods address a more complex version of the problem by relaxing the assumption of mutation independence. This approach is justified by the widespread occurrence of epistasis —- non-additive phenotypic effects of combinations of mutations [217, 233–238]. The presence of epistasis suggests that selection acts on haplotypes, or combinations of mutations, rather than on individual mutations, contributing to the non-linearity of viral evolution and making forecasting more challenging. However, various methods exist to detect epistatic links between mutations [217,234,236]. The identification of these links, and more broadly epistatic networks, holds promise for improving evolutionary forecasting. In fact, it may enable predictions of viral variants before they emerge, as epistatic interactions influencing their phenotypes can be detected early [230]. For example, the tool HELEN [230] forecasts potentially spreading viral variants by identifying dense communities of mutations in networks of epistatic interactions.

Another exciting development in evolutionary forecasting is the application of Large Language Models (LLMs) [239–242]—advanced artificial intelligence systems trained on large amounts of text data to understand, generate, and process natural language. Biological nucleotide or amino acid sequences, including viral genomes, can be conceptualized as words in a genomic "language," with specific syntax (or grammar) defining viable genomes and semantics (or meaning) dictating their phenotypic features. Machine learning models properly trained on large viral genome datasets can potentially forecast viral evolution by identifying mutations or combinations of mutations that maintain syntax while altering semantics, leading to the emergence of viable variants with modified phenotypes. For instance, a recent study [239] applied LLMs to influenza, HIV-1, and SARSCoV-2 protein sequences to accurately predict escape mutations, allowing these viruses to evade hosts' immune responses and potentially giving rise to novel strains with increased infectivity or virulence.

A demonstration of the potential power of genomic-based evolutionary forecasting is the authors' use of the HELEN (Heralding Emerging Lineages in Epistatic Networks) tool [230]. In this study, SARS-CoV-2 genomic data from 16 countries with high sampling coverage were used to retroactively predict the emergence of Variants of Concern (VOCs) Alpha, Beta, Gamma, Delta, and Omicron, and Variants of Interest (VOIs) Lambda, Mu, Theta, Eta, and Kappa. All VOCs were detected early, with forecasting depths ranging from 30 days (Omicron) to 285 days (Delta) prior to the times when the variants reached 1 % prevalence. Four out of five VOIs were also detected 0 to 75 days before they reached a 1 % prevalence.

Footnotes

N/A.

Data statement/ data linking

All of the data employed in the paper is publicly available.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding sources

GC is partially supported from NSF grants CNS 2125246 and DEB 2026797. PS is partially supported from NSF grants OISE-2412914, CCF-2047828 and IIS-2212508.

Acknowledgements

We thank Amanda Bleichrodt for help proofreading and Elienai Joaquin-Damas for help with the references.

Appendices

N/A.

References

[1] Holmdahl I, Buckee C. Wrong but useful — what covid-19 epidemiologic models can and cannot tell us. N Engl J Med 2020;383:303–5. https://doi.org/10.1056/nejmp2016822.
[2] Anderson RM, May RM. Infectious diseases of humans: dynamics and control. 1st ed. Oxford: Oxford University Press; 1991.
[3] Diekmann O, Heesterbeek JAP. Mathematical epidemiology of infectious diseases: model building, analysis and interpretation. 1st ed. West Sussex: Jhon Wiley & Sons Ltd; 2000.
[4] Brauer F, Castillo-Chavez C, Feng Z. Mathematical models in epidemiology. Berlin: Springer; 2019. https://doi.org/10.1007/978-1-4939-9828-9.
[5] Keeling M j, Rohani P. Modeling infectious diseases in humans and animals. Princeton University Press; 2008. https://doi.org/10.2307/j.ctvcm4gk0.
[6] Ofori SK, Schwind JS, Sullivan KL, Chowell G, Cowling BJ, Fung ICH. Age-stratified model to assess health outcomes of COVID-19 vaccination strategies, Ghana. Emerg Infect Dis 2023;29:360–70. https://doi.org/10.3201/eid2902.221098.
[7] Cramer EY, Ray EL, Lopez VK, Bracher J, Brennen A, Castro Rivadeneira AJ, et al. Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States. PNAS 2022;15:1–12. https://doi.org/10.1073/pnas.2113561119. e2113561119.
[8] Roosa K, Lee Y, Luo R, Kirpich A, Rothenberg R, Hyman JM, et al. Real-time forecasts of the COVID-19 epidemic in China from February 5th to February 24th, 2020. Infect Dis Model 2020;5:256–63. https://doi.org/10.1016/j.idm.2020.02.002.
[9] Anastassopoulou C, Russo L, Tsakris A, Siettos C. Data-based analysis, modelling and forecasting of the COVID-19 outbreak. PLoS One 2020;15. https://doi.org/10.1371/journal.pone.0230405.
[10] Friedman J, Liu P, Troeger CE, Carter A, Reiner RC, Barber RM, et al. Predictive performance of international COVID-19 mortality forecasting models. Nat Commun 2021;12. https://doi.org/10.1038/s41467-021-22457-w.
[11] Biggerstaff M, Cauchemez S, Reed C, Gambhir M, Finelli L. Estimates of the reproduction number for seasonal, pandemic, and zoonotic influenza: a systematic review of the literature. BMC Infect Dis 2014;14. https://doi.org/10.1186/1471-2334-14-480.
[12] Funk S, Camacho A, Kucharski AJ, Lowe R, Eggo RM, Edmunds WJ. Assessing the performance of real-time epidemic forecasts: a case study of ebola in the Western area region of sierra leone. PLoS Comput Biol 2019;15:2014–5. https://doi.org/10.1371/journal.pcbi.1006785.
[13] Chowell G, Simonsen L, Viboud C, Kuang Y. Is West Africa approaching a catastrophic phase or is the 2014 Ebola epidemic slowing down? Different models yield different answers for liberia. PLoS Curr 2014. https://doi.org/10.1371/currents.outbreaks.b4690859d91684da963dc40e00f3da81.
[14] Roosa K, Tariq A, Yan P, Hyman JM, Chowell G. Multi-model forecasts of the ongoing Ebola epidemic in the Democratic Republic of Congo, March-October 2019. J R Soc Interface 2020;17. https://doi.org/10.1098/rsif.2020.0447.
[15] Bleichrodt A, Luo R, Kirpich A, Chowell G. Evaluating the forecasting performance of ensemble sub-epidemic frameworks and other time series models for the 2022–2023 mpox epidemic. R Soc Open Sci 2024;11. https://doi.org/10.1098/rsos.240248.
[16] Overton CE, Abbott S, Christie R, Cumming F, Day J, Jones O, et al. Nowcasting the 2022 mpox outbreak in England. PLoS Comput Biol 2023;19. https://doi.org/10.1371/journal.pcbi.1011463.
[17] Brand SPC, Cavallaro M, Cumming F, Turner C, Florence I, Blomquist P, et al. The role of vaccination and public awareness in forecasts of Mpox incidence in the United Kingdom. Nat Commun 2023;14. https://doi.org/10.1038/s41467-023-38816-8.
[18] Kaftan D, Kim HY, Ko C, Howard JS, Dalal P, Yamamoto N, et al. Performance analysis of mathematical methods used to forecast the 2022 New York City Mpox outbreak. J Med Virol 2024;96. https://doi.org/10.1002/jmv.29791.
[19] Jaqaman K, Danuser G. Linking data to models: data regression. Nat Rev Mol Cell Biol 2006;7:813–9. https://doi.org/10.1038/nrm2030.
[20] Wieland FG, Hauber AL, Rosenblatt M, Tönsing C, Timmer J. On structural and practical identifiability. Curr Opin Syst Biol 2021;25:60–9. https://doi.org/10.1016/j.coisb.2021.03.005.
[21] Raue A, Kreutz C, Maiwald T, Bachmann J, Schilling M, Klingmüller U, et al. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. Bioinformatics 2009;25:1923–9. https://doi.org/10.1093/bioinformatics/btp358.
[22] Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis 2020;20:533–4.
[23] Xu B, Kraemer MUG, Gutierrez B, Mekaru S, Sewalk K, Loskill A, et al. Open access epidemiological data from the COVID-19 outbreak. Lancet Infect Dis 2020;20:534. https://doi.org/10.1016/S1473-3099(20)30119-5.
[24] Grantz KH, Meredith HR, Cummings DAT, Metcalf CJE, Grenfell BT, Giles JR, et al. The use of mobile phone data to inform analysis of COVID-19 pandemic epidemiology. Nat Commun 2020;11. https://doi.org/10.1038/s41467-020-18190-5.
[25] Sun K, Chen J, Viboud C. Early epidemiological analysis of the coronavirus disease 2019 outbreak based on crowdsourced data: a population-level observational study. Lancet Digit Health 2020;2:e201–8. https://doi.org/10.1016/S2589-7500(20)30026-1.
[26] Vahedi B, Becker AD, Wesolowski A. Crowdsourcing the landscape of COVID-19 data dashboards amid the pandemic: a call for collaboration. PLOS Digital Health 2022;1:e0000004.
[27] Bertozzi AL, Franco E, Mohler G, Short MB, Sledge D. The challenges of modeling and forecasting the spread of COVID-19. Proc Natl Acad Sci U S A 2020;117:16732–8. https://doi.org/10.1073/pnas.2006520117.
[28] Kermack W0, Mckendrick AG. A contribution to the mathematical theory of epidemics. R Soc 1927;115:700–21.
[29] Hethcote HW. The mathematics of infectious diseases. SIAM Rev 2000;42:599–653.
[30] Wallinga J, Lipsitch M. How generation intervals shape the relationship between growth rates and reproductive numbers. Proc R Soc B: Biol Sci 2007;274:599–604.
[31] Van Den Driessche P, Watmough J. Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission. Math Biosci 2002;180:29–48.
[32] Chowell G, Sattenspiel L, Bansal S, Viboud C. Mathematical models to characterize early epidemic growth: a review. Phys Life Rev 2016;18:66–97. https://doi.org/10.1016/j.plrev.2016.07.005.
[33] Viboud C, Simonsen L, Chowell G. A generalized-growth model to characterize the early ascending phase of infectious disease outbreaks. Epidemics 2016;15:27–37. https://doi.org/10.1016/j.epidem.2016.01.002.
[34] Zhao Musa, Lin Ran, Yang Wang, et al. Estimating the unreported number of novel coronavirus (2019-nCoV) cases in China in the first half of January 2020: a data-driven modelling analysis of the early outbreak. JCM 2020;9:388.
[35] Pellis L, Scarabel F, Stage HB, Overton CE, Chappell LHK, Fearon E, et al. Challenges in control of COVID-19: short doubling time and long delay to effect of interventions. Philosoph Trans R Soc B: Biol Sci 2021;376. https://doi.org/10.1098/rstb.2020.0264.
[36] Roosa K, Lee Y, Luo R, Kirpich A, Rothenberg R, Hyman JM, et al. Real-time forecasts of the COVID-19 epidemic in China from February 5th to February 24th, 2020. Infect Dis Model 2020;5:256–63.
[37] Vynnycky E, White RG. An introduction to infectious diseas modelling. 1st ed. Oxford: Oxford University Press; 2010.
[38] Heffernan JM, Smith RJ, Wahl LM. Perspectives on the basic reproductive ratio. J R Soc Interface 2005;2:281–93. https://doi.org/10.1098/rsif.2005.0042.
[39] Grassly NC, Fraser C. Mathematical models of infectious disease transmission. Nat Rev Microbiol 2008;6:477–87. https://doi.org/10.1038/nrmicro1845.

[40] Ma J. Estimating epidemic exponential growth rate and basic reproduction number. Infect Dis Model 2020;5:129–41. https://doi.org/10.1016/j.idm.2019.12.009.
[41] Tsoularis A, Wallace J. Analysis of logistic growth models. Math Biosci 2002;179:21–55.
[42] Yan P, Chowell G. Quantitative Methods for Investigating Infectious Disease Outbreaks. 1st ed., 70. Springer; 2019. https://doi.org/10.1007/978-3-030-21923-9.
[43] O'Neill PD, Roberts GO. Bayesian inference for partially observed stochastic epidemics. J R Stat Soc Ser A Stat Soc 1999;162:121–9.
[44] Birrell PJ, de Angelis D, Presanis AM. Evidence synthesis for stochastic epidemic models. Stat Sci 2018;33:34–43. https://doi.org/10.1214/17-STS631.
[45] Zhao S, Musa SS, Fu H, He D, Qin J. Simple framework for real-time forecast in a data-limited situation: the Zika virus (ZIKV) outbreaks in Brazil from 2015 to 2016 as an example. Parasit Vect 2019;12. https://doi.org/10.1186/s13071-019-3602-9.
[46] Wei ZL, Wang DF, Sun HY, Yan X. Comparison of a physical model and phenomenological model to forecast groundwater levels in a rainfall-induced deep-seated landslide. J Hydrol (Amst) 2020;586:1–9. https://doi.org/10.1016/j.jhydrol.2020.124894.
[47] Pell B, Kuang Y, Viboud C, Chowell G. Using phenomenological models for forecasting the 2015 Ebola challenge. Epidemics 2018;22:62–70. https://doi.org/10.1016/j.epidem.2016.11.002.
[48] Lega J, Brown HE. Data-driven outbreak forecasting with a simple nonlinear growth model. Epidemics 2016;17:19–26. https://doi.org/10.1016/j.epidem.2016.10.002.
[49] Lasky JR, Hooten MB, Adler PB. What processes must we understand to forecast regional-scale population dynamics?: regional population forecasting. Proc R Soc B: Biol Sci 2020;287. https://doi.org/10.1098/rspb.2020.2219.
[50] Hsieh Y-H, Cheng Y-S. Real-time forecast of multiphase outbreak. Emerg Infect Dis 2006;12:122–7.
[51] Chowell G, Hincapie-Palacio D, Ospina J, Pell B, Tariq A, Dahal S, et al. Using phenomenological models to characterize transmissibility and forecast patterns and final burden of Zika epidemics. PLoS Curr 2016. https://doi.org/10.1371/currents.outbreaks.f14b2217c902f453d9320a43a35b9583.
[52] Shanafelt DW, Jones G, Lima M, Perrings C, Chowell G. Forecasting the 2001 foot-and-mouth disease epidemic in the UK. Ecohealth 2018;15:338–47. https://doi.org/10.1007/s10393-017-1293-2.
[53] Roosa K, Lee Y, Luo R, Kirpich A, Rothenberg R, Hyman JM, et al. Short-term forecasts of the COVID-19 epidemic in Guangdong and Zhejiang, China: February 13–23, 2020. J Clin Med 2020;9. https://doi.org/10.3390/jcm9020596.
[54] Chowell G, Viboud C, Hyman JM, Simonsen L. The Western Africa Ebola virus disease epidemic exhibits both global exponential and local polynomial growth rates. PLoS Curr 2015;7. https://doi.org/10.1371/currents.outbreaks.8b55f4bad99ac5c5db3663e916803261.
[55] Brauer F, Castillo-Chavez C. Mathematical models in population biology and epidemiology. 2nd Edition, 2. Springer; 2012.
[56] Dietz K. The estimation of the basic reproduction number for infectious diseases. Stat Methods Med Res 1993;2:23–41. https://doi.org/10.1177/096228029300200103.
[57] Brauer F, Castillo-Chavez C, Feng Z. Simple compartmental models for disease transmission. Math Models Epidemiol 2019:21–61. https://doi.org/10.1007/978-1-4939-9828-9_2.
[58] Diekmann O, Heesterbeek H, Britton T. Mathematical tools for understanding infectious disease dynamics. Princeton, New Jersey: Princeton University Press; 2013.
[59] Chowell G, Dahal S, Liyanage YR, Tariq A, Tuncer N. Structural identifiability analysis of epidemic models based on differential equations: a tutorial-based primer. J Math Biol 2023;87:1–65. https://doi.org/10.1007/s00285-023-02007-2.
[60] Miao H, Xia X, Perelson AS, Wu H. On identifiability of nonlinear ODE models and applications in viral dynamics. SIAM Rev 2011;53:3–39. https://doi.org/10.1137/090757009.
[61] Massonis G, Banga JR, Villaverde AF. Structural identifiability and observability of compartmental models of the COVID-19 pandemic. Annu Rev Control 2021;51:441–59. https://doi.org/10.1016/j.arcontrol.2020.12.001.
[62] DiStefano III JJ, Cobelli C. On parameter and structural identifiability: nonunique observability/reconstructibility for identifiable systems, other ambiguities, and new definitions. IEEB Trans Autom Control 1980. AC-250018-9286/80/08oCr0830.00.75.
[63] Raue A, Karlsson J, Saccomani MP, Jirstrand M, Timmer J. Comparison of approaches for parameter identifiability analysis of biological systems. Bioinformatics 2014;30:1440–8. https://doi.org/10.1093/bioinformatics/btu006.
[64] Pohjanpalo H. System identifiability based on the power series expansion of the solution. Math Biosci 1978;41:21–33. https://doi.org/10.1016/0025-5564(78)90063-9.
[65] Walter E, Lecourtier Y. Global approaches to identifiability testing for linear and nonlinear state space models. North-Holland Publishing Company; 1982.
[66] Vajda S, Godfrey KR, Rabitz H. Similarity transformation approach to identifiability analysis of nonlinear compartmental models. Math Biosci 1989;93:217–48.
[67] Denis-Vidal L, Joly-Blanchard G, Noiret C. Some effective approaches to check the identifiability of uncontrolled nonlinear systems. Math Comput Simul 2001;57:35–44.
[68] Bellu G, Saccomani MP, Audoly S, D'Angiò L. DAISY: a new software tool to test global identifiability of biological and physiological systems. Comput Methods Programs Biomed 2007;88:52–61. https://doi.org/10.1016/j.cmpb.2007.07.002.
[69] Ljung L, Glad T. On global identifiability for arbitrary model parametrizations. Automatica 1994;30:265–76. https://doi.org/10.1016/0005-1098(94)90029-9.
[70] Balsa-Canto E, Banga JR. AMIGO, a toolbox for advanced model identification in systems biology using global optimization. Bioinformatics 2011;27:2311–3. https://doi.org/10.1093/bioinformatics/btr370.
[71] Meshkat N, Er-zhen Kuo C, DiStefano J. On finding and using identifiable parameter combinations in nonlinear dynamic systems biology models and combos: a novel web implementation. PLoS One 2014;9. https://doi.org/10.1371/journal.pone.0110261.
[72] Hong H, Ovchinnikov A, Pogudin G, Yap C. SIAN: software for structural identifiability analysis of ODE models. Bioinformatics 2019;35:2873–4. https://doi.org/10.1093/bioinformatics/bty1069.
[73] Dong R, Goodbrake C, Harrington HA, Pogudin G. Differential elimination for dynamical models via projections with applications to structural identifiability. SIAM J Appl Algebr Geom 2023;7:1–42.
[74] Saccomani MP, Audoly S, Bellu G, D'Angiò L. Examples of testing global identifiability of biological and biomedical models with the DAISY software. Comput Biol Med 2010;40:402–7. https://doi.org/10.1016/j.compbiomed.2010.02.004.
[75] Cobelli C, DiStefano JJ. Parameter and structural identifiability concepts and ambiguities: a critical review and analysis. Am J Physiol-Regul, Integr Compar Physiol 1980;239. https://doi.org/10.1152/ajpregu.1980.239.1.R7.
[76] Tuncer N, Le TT. Structural and practical identifiability analysis of outbreak models. Math Biosci 2018;299:1–18. https://doi.org/10.1016/j.mbs.2018.02.004.
[77] Villaverde AF, Barreiro A, Papachristodoulou A. Structural identifiability of dynamic systems biology models. PLoS Comput Biol 2016;12. https://doi.org/10.1371/journal.pcbi.1005153.
[78] Chowell G. Fitting dynamic models to epidemic outbreaks with quantified uncertainty: a primer for parameter uncertainty, identifiability, and forecasts. Infect Dis Model 2017;2:379–98. https://doi.org/10.1016/j.idm.2017.08.001.
[79] Chowell G, Nishiura H, Bettencourt LMA. Comparative estimation of the reproduction number for pandemic influenza from daily case notification data. J R Soc Interface 2007;4:154–66. https://doi.org/10.1098/rsif.2006.0161.
[80] Smirnova A, deCamp L, Chowell G. Forecasting epidemics through nonparametric estimation of time-dependent transmission rates using the SEIR model. Bull Math Biol 2019;81:4343–65. https://doi.org/10.1007/s11538-017-0284-3.
[81] Earn DJD, Park SW, Bolker BM. Fitting Epidemic models to data: a tutorial in memory of Fred Brauer. Bull Math Biol 2024;86. https://doi.org/10.1007/s11538-024-01326-9.
[82] Chowell G, Bleichrodt A, Luo R. Parameter estimation and forecasting with quantified uncertainty for ordinary differential equation models using QuantDiffForecast: a MATLAB toolbox and tutorial. Stat Med 2024;43:1826–48.

[83] Bracher J, Ray EL, Gneiting T, Reich NG. Evaluating epidemic forecasts in an interval format. PLoS Comput Biol 2021;17:1–15. https://doi.org/10.1371/JOURNAL.PCBI.1008618.
[84] Lloyd-Smith JO. Maximum likelihood estimation of the negative binomial dispersion parameter for highly overdispersed data, with applications to infectious diseases. PLoS One 2007;2. https://doi.org/10.1371/journal.pone.0000180.
[85] Karami H, Bleichrodt A, Luo R, Chowell G. BayesianFitForecast: a user-friendly R toolbox for parameter estimation and forecasting with ordinary differential equations. Under Rev 2024.
[86] Roda WC. Bayesian inference for dynamical systems. Infect Dis Model 2020;5:221–32. https://doi.org/10.1016/j.idm.2019.12.007.
[87] Linden NJ, Kramer B, Rangamani P. Bayesian parameter estimation for dynamical models in systems biology. PLoS Comput Biol 2022;18. https://doi.org/10.1371/journal.pcbi.1010651.
[88] Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, et al. Stan: a probabilistic programming language. J Stat Softw 2017;76. https://doi.org/10.18637/jss.v076.i01.
[89] King AA, Nguyen D, Ionides EL. Statistical inference for partially observed markov processes via the R package pomp. J Stat Softw 2016;69:1–43. https://doi.org/10.18637/jss.v069.i12.
[90] Guedj J, Thiébaut R, Commenges D. Practical identifiability of HIV dynamics models. Bull Math Biol 2007;69:2493–513. https://doi.org/10.1007/s11538-007-9228-7.
[91] Roosa K, Chowell G. Assessing parameter identifiability in compartmental dynamic models using a computational approach: application to infectious disease transmission models. Theor Biol Med Model 2019;16. https://doi.org/10.1186/s12976-018-0097-6.
[92] Banks HT, Tran HT. Mathematical and experimental modeling of physical and biological processes. 1st ed. Boca Ratón, Florida: CRC Press, Taylor & Francis Group, Chapman & Hall Books; 2009.
[93] Eisenberg MC, Robertson SL, Tien JH. Identifiability and estimation of multiple transmission pathways in cholera and waterborne disease. J Theor Biol 2013;324:84–102. https://doi.org/10.1016/j.jtbi.2012.12.021.
[94] Banks HT, Shuhua H, Thompson WC. Modeling and inverse problems in the presence of uncertainty. 1st ed. New York: Chapman and Hall/CRC; 2014.
[95] Simpson MJ, Maclaren OJ. Profile-wise analysis: a profile likelihood-based workflow for identifiability analysis, estimation, and prediction with mechanistic mathematical models. PLoS Comput Biol 2023;19. https://doi.org/10.1371/journal.pcbi.1011515.
[96] Eisenberg MC, Hayashi MAL. Determining identifiable parameter combinations using subset profiling. Math Biosci 2014;256:116–26. https://doi.org/10.1016/j.mbs.2014.08.008.
[97] Tuncer N, Timsina A, Nuno M, Chowell G, Martcheva M. Parameter identifiability and optimal control of an SARS-CoV-2 model early in the pandemic. J Biol Dyn 2022;16:412–38. https://doi.org/10.1080/17513758.2022.2078899.
[98] Nemeth L, Tuncer N, Martcheva M. Structural and practical identifiability analysis of a multiscale immuno-epidemiological model. editors. In: Tuncer N, Martcheva M, Prosper O, Childs L, editors. Computational and mathematical population dynamics. World Scientific Pub Co Inc; 2023. p. 169–201.
[99] Zhang S, Ponce J, Zhang Z, Lin G, Karniadakis G. An integrated framework for building trustworthy data-driven epidemiological models: application to the COVID-19 outbreak in New York City. PLoS Comput Biol 2021;17. https://doi.org/10.1371/journal.pcbi.1009334.
[100] Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. Bayesian data analysis. 3rd, 1. Chapman and Hall/CRC; 2015.
[101] Bell BP, Damon IK, Jernigan DB, Kenyon TA, O'Connor JP, Tappero JW. Overview, control strategies, and lessons learned in the CDC response to the 2014–2016 Ebola epidemic. MMWR 2016;65(Suppl):1–8.
[102] Riley S, Fraser C, Donnelly CA, Ghani AC, Abu-Raddad LJ, Hedley AJ, et al. Transmission dynamics of the etiological agent of SARS in Hong Kong: impact of public health interventions. Science (1979) 2003;300:1961–6. https://doi.org/10.1126/science.1084146.
[103] Chowell G, Fenimore PW, Castillo-Garsow MA, Castillo-Chavez C. SARS outbreaks in Ontario, Hong Kong and Singapore: the role of diagnosis and isolation as a control mechanism. J Theor Biol 2003;224:1–8. https://doi.org/10.1016/S0022-5193(03)00228-5.
[104] Melikechi O, Young AL, Tang T, Bowman T, Dunson D, Johndrow J. Limits of epidemic prediction using SIR models. J Math Biol 2022;85. https://doi.org/10.1007/s00285-022-01804-5.
[105] US Department of Health and Human Services. An HHS retrospective on the 2009 H1N1 influenza pandemic to advance. all hazards preparedness. Washington, DC: 2012.
[106] Meltzer MI, Atkins CY, Santibanez S, Knust B, Petersen BW, Ervin ED, et al. Estimating the future number of cases in the Ebola epidemic–Liberia and Sierra Leone, 2014-2015. MMWR 2014;63(Suppl):1–14.
[107] Chowell G, Nishiura H. Transmission dynamics and control of Ebola Virus Disease (EVD): A review. BMC Med 2014;12:1–17. https://doi.org/10.1186/s12916-014-0196-0.
[108] Viboud C, Sun K, Gaffey R, Ajelli M, Fumanelli L, Merler S, et al. The RAPIDD Ebola forecasting challenge: Synthesis and lessons learnt. Epidemics 2018;22:13–21. https://doi.org/10.1016/j.epidem.2017.08.002.
[109] Das P, Igoe M, Lacy A, Farthing T, Timsina A, Lanzas C, et al. Modeling county level COVID-19 transmission in the greater St. Louis area: challenges of uncertainty and identifiability when fitting mechanistic models to time-varying processes. Math Biosci 2024;371. https://doi.org/10.1016/j.mbs.2024.109181.
[110] García-Basteiro A, Alvarez-Dardet C, Arenas A, Bengoa R, Borrell C, Del Val M, et al. The need for an independent evaluation of the COVID-19 response in Spain. Lancet 2020;396:529–30. https://doi.org/10.1016/S0140-6736(20)31713-X.
[111] Yan P, Chowell G. Modeling sub-exponential epidemic growth dynamics through unobserved individual heterogeneity: a frailty model approach. Math Biosci Eng 2024;21:7278–96. https://doi.org/10.3934/mbe.2024321.
[112] Chowell G, Viboud C, Simonsen L, Moghadas SM. Characterizing the reproduction number of epidemics with early subexponential growth dynamics. J R Soc Interface 2016;13. https://doi.org/10.1098/rsif.2016.0659.
[113] Zhang Y, Long M, Chen K, Xing L, Jin R, Jordan MI, et al. Skilful nowcasting of extreme precipitation with NowcastNet. Nature 2023;619:526–32. https://doi.org/10.1038/s41586-023-06184-4.
[114] Bi K, Xie L, Zhang H, Chen X, Gu X, Tian Q. Accurate medium-range global weather forecasting with 3D neural networks. Nature 2023;619:533–8. https://doi.org/10.1038/s41586-023-06185-3.
[115] Li L., Carver R., Lopez-Gomez I., Sha F., Anderson J. Generative emulation of weather forecast ensembles with diffusion models. vol. 10. 2024.
[116] Lam R, Sánchez-González Á, Willson M, Wirnsberger P, Fortunato M, Alet F, et al. Learning skillful medium-range global weather forecasting. Science (1979) 2023;382:1416–21.
[117] Rustam F, Reshi AA, Mehmood A, Ullah S, On BW, Aslam W, et al. COVID-19 future forecasting using supervised machine learning models. IEEE Access 2020;8:101489–99. https://doi.org/10.1109/ACCESS.2020.2997311.
[118] Rodríguez A., Tabassum A., Cui J., Xie J., Ho J., Agarwal P., et al. DeepCOVID: an operational deep learning-driven framework for explainable real-time COVID-19 forecasting. 2021.
[119] Petropoulos F, Makridakis S. Forecasting the novel coronavirus COVID-19. PLoS One 2020;15. https://doi.org/10.1371/journal.pone.0231236.
[120] Lim B, Zohren S. Time-series forecasting with deep learning: a survey. Philosoph Trans R Soc A: Math, Phys Eng Sci 2021;379. https://doi.org/10.1098/rsta.2020.0209.
[121] Liang Y, Wen H, Nie Y, Jiang Y, Jin M, Song D, et al. Foundation models for time series analysis: a tutorial and survey. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery; 2024. p. 6555–65. https://doi.org/10.1145/3637528.3671451.
[122] Chimmula VKR, Zhang L. Time series forecasting of COVID-19 transmission in Canada using LSTM networks. Chaos Solitons Fract 2020;135. https://doi.org/10.1016/j.chaos.2020.109864.
[123] Shahid F, Zameer A, Muneeb M. Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM. Chaos Solitons Fract 2020;140. https://doi.org/10.1016/j.chaos.2020.110212.

[124] Dandekar R, Barbastathis G. Quantifying the effect of quarantine control in Covid-19 infectious spread using machine learning. MedRxiv 2020;24:102372. https://doi.org/10.1101/2020.04.03.20052084.
[125] Madden WG, Jin W, Lopman B, Zufle A, Dalziel B, Jessica C, et al. Neural networks for endemic measles dynamics: Comparative analysis and integration with mechanistic models. MedRxiv 2024:1–18. https://doi.org/10.1101/2024.05.28.24307979.
[126] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 2019;1:206–2015. https://doi.org/10.1038/s42256-019-0048-x.
[127] Roscher R, Bohn B, Duarte MF, Garcke J. Explainable machine learning for scientific insights and discoveries. IEEE Access 2020;8:42200–16. https://doi.org/10.1109/ACCESS.2020.2976199.
[128] Karniadakis GE, Kevrekidis IG, Lu L, Perdikaris P, Wang S, Yang L. Physics-informed machine learning. Nat Rev Phys 2021;3:422–40. https://doi.org/10.1038/s42254-021-00314-5.
[129] Chen RTQ, Rubanova Y, Bettencourt J, Duvenaud D. Neural ordinary differential equations. Neural Information Processing Systems. 2018. p. 1–18. 32nd Conference, https://doi.org/10.48550/arXiv.1806.07366.
[130] Willard J, Jia X, Xu S, Steinbach M, Kumar V. Integrating physics-based modeling with machine learning: a survey. ArXiv 2020;1:1–34. https://doi.org/10.1145/1122445.1122456.
[131] Noordijk B, Garcia Gomez ML, ten Tusscher KHWJ, de Ridder D, van Dijk ADJ, Smith RW. The rise of scientific machine learning: a perspective on combining mechanistic modelling with machine learning for systems biology. Front Syst Biol 2024;4. https://doi.org/10.3389/fsysb.2024.1407994.
[132] Rackauckas C., Ma Y., Martensen J., Warner C., Zubov K., Supekar R., et al. Universal differential equations for scientific machine learning. PNAS 2020;XXX:1–6. https://doi.org/10.1073/pnas.XXXXXXXXXX.
[133] Procopio A, Cesarelli G, Donisi L, Merola A, Amato F, Cosentino C. Combined mechanistic modeling and machine-learning approaches in systems biology – A systematic literature review. Comput Methods Programs Biomed 2023;240. https://doi.org/10.1016/j.cmpb.2023.107681.
[134] Armstrong GL, MacCannell DR, Taylor J, Carleton HA, Neuhaus EB, Bradbury RS, et al. Pathogen genomics in public health. N Engl J Med 2019;381:2569–80.
[135] Black A, MacCannell DR, Sibley TR, Bedford T. Ten recommendations for supporting open pathogen genomic analysis in public health. Nat Med 2020;26:832–41. https://doi.org/10.1038/s41591-020-0935-z.
[136] Knyazev S, Chhugani K, Sarwal V, Ayyala R, Singh H, Karthikeyan S, et al. Unlocking capacities of genomics for the COVID-19 response and future pandemics. Nat Methods 2022;19:374–80.
[137] Jombart T, Cori A, Didelot X, Cauchemez S, Fraser C, Ferguson N. Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. PLoS Comput Biol 2014;10. https://doi.org/10.1371/journal.pcbi.1003457.
[138] Campbell F, Didelot X, Fitzjohn R, Ferguson N, Cori A, Jombart T. Outbreaker2: a modular platform for outbreak reconstruction. BMC Bioinform 2018;19:1–8.
[139] Jombart T, Eggo RM, Dodd PJ, Balloux F. Reconstructing disease outbreaks from genetic data: a graph approach. Heredity (Edinb) 2011;106:383–90. https://doi.org/10.1038/hdy.2010.78.
[140] De Maio N, Wu C-H, Wilson DJ. SCOTTI: efficient reconstruction of transmission within outbreaks with the structured coalescent. PLoS Comput Biol 2016;12.
[141] Skums P, Mohebbi F, Tsyvina V, Baykal PI, Nemira A, Ramachandran S, et al. SOPHIE: viral outbreak investigation and transmission history reconstruction in a joint phylogenetic and network theory framework. Cell Syst 2022;13:844–56. https://doi.org/10.1016/j.cels.2022.07.005. e4.
[142] Klinkenberg D, Backer JA, Didelot X, Colijn C, Wallinga J. Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. PLoS Comput Biol 2017;13.
[143] Worby CJ, O'Neill PD, Kypraios T, Robotham JV, De Angelis D, Cartwright EJP, et al. Reconstructing transmission trees for communicable diseases using densely sampled genetic data. Ann Appl Stat 2016;10:395.
[144] De Maio N, Worby CJ, Wilson DJ, Stoesser N. Bayesian reconstruction of transmission within outbreaks using genomic variants. PLoS Comput Biol 2018;14.e1006117–e1006117.
[145] Wymant C, Hall M, Ratmann O, Bonsall D, Golubchik T, De Cesare M, et al. PHYLOSCANNER: inferring transmission from within- and between-host pathogen genetic diversity. Mol Biol Evol 2018;35:719–33. https://doi.org/10.1093/molbev/msx304.
[146] de Bernardi Schneider A, Ford CT, Hostager R, Williams J, Cioce M, Catalyürek ÜV, et al. StrainHub: a phylogenetic tool to construct pathogen transmission networks. Bioinformatics 2020;36:945–7.
[147] Didelot X, Fraser C, Gardy J, Colijn C, Malik H. Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. Mol Biol Evol 2017;34:997–1007. https://doi.org/10.1093/molbev/msw275.
[148] Didelot X, Gardy J, Colijn C. Bayesian inference of infectious disease transmission from whole-genome sequence data. Mol Biol Evol 2014;31:1869–79. https://doi.org/10.1093/molbev/msu121.
[149] Carson J, Keeling M, Wyllie D, Ribeca P, Didelot X. Inference of infectious disease transmission through a relaxed bottleneck using multiple genomes Per Host. Mol Biol Evol 2024;41. https://doi.org/10.1093/molbev/msad288.
[150] Hall MD, Colijn C. Transmission trees on a known pathogen phylogeny: enumeration and sampling. Mol Biol Evol 2019;36:1333–43. https://doi.org/10.1093/molbev/msz058.
[151] Sledzieski S, Zhang C, Mandoiu I, Bansal MS. TreeFix-TP: phylogenetic error-correction for infectious disease transmission network inference. Pac Symp Biocomput 2021:119–30. https://doi.org/10.1142/9789811232701_0012open_in_new.
[152] Skums P, Zelikovsky A, Singh R, Gussler W, Dimitrova Z, Knyazev S, et al. QUENTIN: reconstruction of disease transmissions from viral quasispecies genomic data. Bioinformatics 2018;34:163–70.
[153] Glebova O, Knyazev S, Melnyk A, Artyomenko A, Khudyakov Y, Zelikovsky A, et al. Inference of genetic relatedness between viral quasispecies from sequencing data. BMC Genomics 2017;18:918. https://doi.org/10.1186/s12864-017-4274-5.
[154] Kosakovsky Pond SL, Weaver S, Leigh Brown AJ, Wertheim JO. HIV-TRACE (TRAnsmission Cluster Engine): a tool for large scale molecular epidemiology of HIV-1 and other rapidly evolving pathogens. Mol Biol Evol 2018;35:1812–9.
[155] Longmire AG, Sims S, Rytsareva I, Campo DS, Skums P, Dimitrova Z, et al. GHOST: global hepatitis outbreak and surveillance technology. BMC Genomics 2017;18:916.
[156] Campbell EM, Boyles A, Shankar A, Kim J, Knyazev S, Cintron R, et al. MicrobeTrace: retooling molecular epidemiology for rapid public health response. PLoS Comput Biol 2021;17. https://doi.org/10.1371/journal.pcbi.1009300.
[157] Sashittal P, El-Kebir M. SharpTNI: Counting and Sampling Parsimonious Transmission Networks under a Weak Bottleneck. BioRxiv. 2019. p. 1–12. https://doi.org/10.1101/842237.
[158] Sashittal P, El-Kebir M. Sampling and summarizing transmission trees with multi-strain infections. Bioinformatics 2020;36:I362–70. https://doi.org/10.1093/BIOINFORMATICS/BTAA438.
[159] Dhar S, Zhang C, Mandoiu II, Bansal MS. TNet: transmission network inference using within-host strain diversity and its application to geographical tracking of COVID-19 spread. IEEE/ACM Trans Comput Biol Bioinform 2022;19:230–42. https://doi.org/10.1109/TCBB.2021.3096455.
[160] Ke Z, Vikalo H. Graph-based reconstruction and analysis of disease transmission networks using viral genomic data. J Comput Biol 2022. https://doi.org/10.1101/2022.07.28.501873.
[161] Kuzmin K, Schmidt H, Snir S, Raphael B, Skums P, Raphael B. Outbreaks, metastases and homomorphisms: phylogenetic inference of migration histories of heterogeneous populations under evolutionary and structural constraints. Nat Commun 2024;00:1–33. https://doi.org/10.21203/rs.3.rs-5040045/v1.
[162] Ypma RJF, van Ballegooijen WM, Wallinga J. Relating phylogenetic trees to transmission trees of infectious disease outbreaks. Genetics 2013;195:1055–62. https://doi.org/10.1534/genetics.113.154856.
[163] Mollentze N, Nel LH, Townsend S, le Roux K, Hampson K, Haydon DT, et al. A Bayesian approach for inferring the dynamics of partially observed endemic infectious diseases from space-time-genetic data. Proc R Soc B: Biol Sci 2014;281. https://doi.org/10.1098/rspb.2013.3251.
[164] Morelli MJ, Thébaud G, Chadœuf J, King DP, Haydon DT. A Bayesian inference framework to reconstruct transmission trees using epidemiological and genetic data. PLoS Comput Biol 2012;8. https://doi.org/10.1371/journal.pcbi.1002768.

[165] Cottam EM, Thébaud G, Wadsworth J, Gloster J, Mansley L, Paton DJ, et al. Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus. Proc R Soc B: Biol Sci 2008;275:887–95. https://doi.org/10.1098/rspb.2007.1442.
[166] Campbell F, Cori A, Ferguson N, Jombart T. Bayesian inference of transmission chains using timing of symptoms, pathogen genomes and contact data. PLoS Comput Biol 2019;15. https://doi.org/10.1371/journal.pcbi.1006930.
[167] Hall M, Woolhouse M, Rambaut A. Epidemic reconstruction in a phylogenetics framework: transmission trees as partitions of the node set. PLoS Comput Biol 2015;11. https://doi.org/10.1371/journal.pcbi.1004613.
[168] Rosset S. Efficient inference on known phylogenetic trees using poisson regression. Bioinformatics 2006;23:142–7. https://doi.org/10.1093/bioinformatics/btl306.
[169] Campo DS, Xia G-L, Dimitrova Z, Lin Y, Forbi JC, Ganova-Raeva L, et al. Accurate genetic detection of hepatitis C virus transmissions in outbreak settings. J Infect Dis 2016;213:957–65.
[170] Wertheim JO, Kosakovsky Pond SL, Forgione LA, Mehta SR, Murrell B, Shah S, et al. Social and genetic networks of HIV-1 transmission in New York City. PLoS Pathog 2017;13. https://doi.org/10.1371/journal.ppat.1006000.
[171] Ratmann O, Grabowski MK, Hall M, Golubchik T, Wymant C, Abeler-Dörner L, et al. Inferring HIV-1 transmission networks and sources of epidemic spread in Africa with deep-sequence phylogenetic analysis. Nat Commun 2019;10. https://doi.org/10.1038/s41467-019-09139-4.
[172] Zhang Y, Wymant C, Laeyendecker O, Grabowski MK, Hall M, Hudelson S, et al. Evaluation of phylogenetic methods for inferring the direction of human immunodeficiency virus (HIV) transmission: HIV prevention trials network (HPTN) 052. Clin Infect Dis 2021;72:30–7. https://doi.org/10.1093/cid/ciz1247.
[173] Ramachandran S, Thai H, Forbi JC, Galang RR, Dimitrova Z, liang Xia G, et al. A large HCV transmission network enabled a fast-growing HIV outbreak in rural Indiana, 2015. EBioMedicine 2018;37:374–81. https://doi.org/10.1016/j.ebiom.2018.10.007.
[174] Knyazev S, Hughes L, Skums P, Zelikovsky A. Epidemiological data analysis of viral quasispecies in the next-generation sequencing era. Brief Bioinform 2020;22:96–108. https://doi.org/10.1093/bib/bbaa101.
[175] Campbell EM, Jia H, Shankar A, Hanson D, Luo W, Masciotra S, et al. Detailed transmission network analysis of a large opiate-driven outbreak of HIV infection in the United States. J Infect Dis 2017;216:1053–62. https://doi.org/10.1093/infdis/jix307. Oxford University Press.
[176] Ragonnet-Cronin M, Hu YW, Morris SR, Sheng Z, Poortinga K, Wertheim JO. HIV transmission networks among transgender women in Los Angeles County, CA, USA: a phylogenetic analysis of surveillance data. Lancet HIV 2019;6:e164–72.
[177] Wertheim JO, Leigh Brown AJ, Hepler NL, Mehta SR, Richman DD, Smith DM, et al. The global transmission network of HIV-1. J Infect Dis 2014;209:304–13. https://doi.org/10.1093/infdis/jit524.
[178] Kong S, Sánchez-Pacheco SJ, Murphy RW. On the use of median-joining networks in evolutionary biology. Cladistics 2016;32:691–9.
[179] Fischer GE, Schaefer MK, Labus BJ, Sands L, Rowley P, Azzam IA, et al. Hepatitis C virus infections from unsafe injection practices at an endoscopy clinic in Las Vegas, Nevada, 2007-2008. Clin Infect Dis 2010;51:267–73. https://doi.org/10.1086/653937.
[180] Romero-Severson EO, Bulla I, Leitner T. Phylogenetically resolving epidemiologic linkage. Proc Natl Acad Sci U S A 2016;113:2690–5. https://doi.org/10.1073/pnas.1522930113.
[181] Leitner T. Phylogenetics in HIV transmission: taking within-host diversity into account. Curr Opin HIV AIDS 2019;14:181–7. https://doi.org/10.1097/COH.0000000000000536.
[182] El-Kebir M, Satas G, Raphael BJ. Inferring parsimonious migration histories for metastatic cancers. Nat Genet 2018;50:718–26. https://doi.org/10.1038/s41588-018-0106-z.
[183] Gusfield D. Integer linear programming in computational and systems biology: an entry-level text and course. Cambridge University Press; 2019.
[184] Nadeau SA, Vaughan TG, Scire J, Huisman JS, Stadler T. The origin and early spread of SARS-CoV-2 in Europe. Proc Natl Acad Sci 2021;118:e2012008118. https://doi.org/10.1073/pnas.2012008118/-/DCSupplemental.
[185] Switzer WM, Shankar A, Jia H, Knyazev S, Ambrosio F, Kelly R, et al. High HIV diversity, recombination, and superinfection revealed in a large outbreak among persons who inject drugs in Kentucky and Ohio, USA. Virus Evol 2024;10. https://doi.org/10.1093/ve/veae015.
[186] Hughes GJ, Fearnhill E, Dunn D, Lycett SJ, Rambaut A, Leigh Brown AJ. Molecular phylodynamics of the heterosexual HIV epidemic in the United Kingdom. PLoS Pathog 2009;5. https://doi.org/10.1371/journal.ppat.1000590.
[187] Liljeros F, Edling CR, Nunes Amaral LA, Stanley HE, Aberg Y. The web of human sexual contacts. Nature 2001;411:907–8. https://doi.org/10.1038/35082140.
[188] Romano CM, de Carvalho-Mello IMVG, Jamal LF, de Melo FL, Iamarino A, Motoki M, et al. Social networks shape the transmission dynamics of hepatitis C virus. PLoS One 2010;5:1–9. https://doi.org/10.1371/journal.pone.0011170.
[189] Newman MEJ. Networks: an introduction. 1st ed. Oxford, New York: Oxford University Press; 2010.
[190] Conrad C, Bradley HM, Broz D, Buddha S, Chapman1 EL, Galang RR, et al. Community outbreak of HIV infection linked to injection drug use of oxymorphone — Indiana, 2015. Morb Mortal Weekly Rep (MMWR) 2015:443–4. https://www.cdc.gov/mmwR/preview/mmwrhtml/mm6416a4.htm. accessed August 7, 2024.
[191] Peters PJ, Pontones P, Hoover KW, Patel MR, Galang RR, Shields J, et al. HIV Infection linked to injection use of oxymorphone in Indiana, 2014–2015. N Engl J Med 2016;375:229–39. https://doi.org/10.1056/nejmoa1515195.
[192] Suryaprasad AG, White JZ, Xu F, Eichler BA, Hamilton J, Patel A, et al. Emerging epidemic of hepatitis C virus infections among young nonurban persons who inject drugs in the United States, 2006-2012. Clin Infect Dis 2014;59:1411–9. https://doi.org/10.1093/cid/ciu643.
[193] Zibbell JE, Iqbal K, Patel RC, Suryaprasad A, Sanders KJ, Moore-Moravian L, et al. Increases in Hepatitis C virus infection related to injection drug use among persons aged ≤30 Years — Kentucky, Tennessee, Virginia, and West Virginia, 2006–2012. MMWR Morb Mortal Wkly Rep 2015;8:453–8. https://doi.org/10.3886/ICPSR25221.v9.
[194] Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 2014;30:1312–3.
[195] Sagulenko P, Puller V, Neher RA. TreeTime: maximum-likelihood phylodynamic analysis. Virus Evol 2018;4. vex042–vex042.
[196] Liu QH, Ajelli M, Aleta A, Merler S, Moreno Y, Vespignani A. Measurability of the epidemic reproduction number in data-driven contact networks. Proc Natl Acad Sci U S A 2018;115:12680–5. https://doi.org/10.1073/pnas.1811115115.
[197] Stadler T. Package "TreePar": estimating birth and death rates based on phylogenies 2015.
[198] Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol 2007;7:214.
[199] Featherstone LA, Zhang JM, Vaughan TG, Duchene S. Epidemiological inference from pathogen genomes: a review of phylodynamic models and applications. Virus Evol 2022;8. https://doi.org/10.1093/ve/veac045.
[200] Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 1981;17:368–76. https://doi.org/10.1007/BF01734359.
[201] Tsyvina V, Zelikovsky A, Snir S, Skums P. Inference of mutability landscapes of tumors from single cell sequencing data. PLoS Comput Biol 2020;16. https://doi.org/10.1371/journal.pcbi.1008454.
[202] Rasmussen DA, Stadler T. Coupling adaptive molecular evolution to phylodynamics using fitness-dependent birth-death models. Elife 2019;8:e45562. https://doi.org/10.7554/eLife.45562.001.
[203] Stadler T. On incomplete sampling under birth-death models and connections to the sampling-based coalescent. J Theor Biol 2009;261:58–66. https://doi.org/10.1016/j.jtbi.2009.07.018.
[204] Pybus OG, Rambaut A, Harvey PH. An integrated framework for the inference of viral population history from reconstructed genealogies. Genet Soc Am 2000;155:1429–37.
[205] Skums P, Tsyvina V, Zelikovsky A. Inference of clonal selection in cancer populations using single-cell sequencing data. Bioinformatics 2019;35:i398–407.
[206] Kühnert D, Stadler T, Vaughan TG, Drummond AJ. Phylodynamics with migration: a computational framework to quantify population structure from genomic data. Mol Biol Evol 2016;33:2102–16. https://doi.org/10.1093/molbev/msw064.
[207] Karcher MD, Palacios JA, Lan S, Minin VN. Phylodyn: an R package for phylodynamic simulation and inference. Mol Ecol Resour 2017;17:96–100. https://doi.org/10.1111/1755-0998.12630. Blackwell Publishing Ltd.

[208] Höhna S, Landis MJ, Heath TA, Boussau B, Lartillot N, Moore BR, et al. RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. Syst Biol 2016;65:726–36. https://doi.org/10.1093/sysbio/syw021.
[209] Lemey P, Rambaut A, Bedford T, Faria N, Bielejec F, Baele G, et al. Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza H3N2. PLoS Pathog 2014;10. https://doi.org/10.1371/journal.ppat.1003932.
[210] Nemira A, Adeniyi AE, Gasich EL, Bulda KY, Valentovich LN, Krasko AG, et al. SARS-CoV-2 transmission dynamics in Belarus revealed by genomic and incidence data analysis. Commun Med 2021;1:1–16. https://doi.org/10.1101/2021.04.13.21255404.
[211] Gankin Y, Nemira A, Koniukhovskii V, Chowell G, Weppelmann TA, Skums P, et al. Investigating the first stage of the COVID-19 pandemic in Ukraine using epidemiological and genomic data. Infect Genet Evol 2021;95. https://doi.org/10.1016/j.meegid.2021.105087.
[212] Kirpich A, Shishkin A, Weppelmann TA, Tchernov AP, Skums P, Gankin Y. Excess mortality in Belarus during the COVID-19 pandemic as the case study of a country with limited non-pharmaceutical interventions and limited reporting. Sci Rep 2022;12. https://doi.org/10.1038/s41598-022-09345-z.
[213] Geoghegan JL, Ren X, Storey M, Hadfield J, Jelley L, Jefferies S, et al. Genomic epidemiology reveals transmission patterns and dynamics of SARS-CoV-2 in Aotearoa New Zealand. Nat Commun 2020;11. https://doi.org/10.1038/s41467-020-20235-8.
[214] Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. The proximal origin of SARS-CoV-2. Nat Med 2020.
[215] Stadler T, Kühnert D, Bonhoeffer S, Drummond AJ. Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). Proc Natl Acad Sci U S A 2013;110:228–33. https://doi.org/10.1073/pnas.1207965110.
[216] Vaughan TG, Leventhal GE, Rasmussen DA, Drummond AJ, Welch D, Stadler T, et al. Estimating epidemic incidence and prevalence from genomic data. Mol Biol Evol 2019;36:1804–16. https://doi.org/10.1093/molbev/msz106.
[217] Rodriguez-Rivas J, Croce G, Muscat M, Weigt M. Epistatic models predict mutable sites in SARS-CoV-2 proteins and epitopes. Proc Natl Acad Sci 2022;119. e2113118119–e2113118119.
[218] Maher MC, Bartha I, Weaver S, Di Iulio J, Ferri E, Soriaga L, et al. Predicting the mutational drivers of future SARS-CoV-2 variants of concern. Sci Transl Med 2022;14. eabk3445–eabk3445.
[219] Icer Baykal PB, Lara J, Khudyakov Y, Zelikovsky A, Skums P. Quantitative differences between intra-host HCV populations from persons with recently established and persistent infections. Virus Evol 2021;7. veaa103–veaa103.
[220] Lässig M, Mustonen V, Walczak AM. Predicting evolution. Nat Ecol Evol 2017;1:1–9.
[221] Łuksza M, Lässig M. A predictive fitness model for influenza. Nature 2014;507:57–61. https://doi.org/10.1038/nature13087.
[222] Davies NG, Abbott S, Barnard RC, Jarvis CI, Kucharski AJ, Munday JD, et al. Estimated transmissibility and impact of SARS-CoV-2 lineage B. 1.1. 7 in England. Science (1979) 2021;372. eabg3055–eabg3055.
[223] Starr TN, Greaney AJ, Hilton SK, Ellis D, Crawford KHD, Dingens AS, et al. Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. Cell 2020;182:1295–310.
[224] Wang P, Nair MS, Liu L, Iketani S, Luo Y, Guo Y, et al. Antibody resistance of SARS-CoV-2 variants B. 1.351 and B. 1.1. 7. Nature 2021;593:130–5.
[225] Planas D, Veyer D, Baidaliuk A, Staropoli I, Guivel-Benhassine F, Rajah MM, et al. Reduced sensitivity of SARS-CoV-2 variant Delta to antibody neutralization. Nature 2021;596:276–80.
[226] Garcia-Beltran WF, Lam EC, Denis KS, Nitido AD, Garcia ZH, Hauser BM, et al. Multiple SARS-CoV-2 variants escape neutralization by vaccine-induced humoral immunity. Cell 2021;184:2372–83.
[227] Barrat-Charlaix P, Huddleston J, Bedford T, Neher RA. Limited predictability of amino acid substitutions in seasonal influenza viruses. Mol Biol Evol 2021;38:2767–77.
[228] Ahmed SF, Quadeer AA, McKay MR. COVIDep: a web-based platform for real-time reporting of vaccine target recommendations for SARS-CoV-2. Nat Protoc 2020;15:2141–2.
[229] Bai C, Wang J, Chen G, Zhang H, An K, Xu P, et al. Predicting mutational effects on receptor binding of the spike protein of SARS-CoV-2 variants. J Am Chem Soc 2021;143:17646–54.
[230] Mohebbi F, Zelikovsky A, Mangul S, Chowell G, Skums P. Early detection of emerging viral variants through analysis of community structure of coordinated substitution networks. Nat Commun 2024;15. https://doi.org/10.1038/s41467-024-47304-6.
[231] Obermeyer F, Jankowiak M, Barkas N, Schaffner SF, Pyle JD, Yurkovetskiy L, et al. Analysis of 6.4 million SARS-CoV-2 genomes identifies mutations associated with fitness. Science (1979) 2022;376:1327–32.
[232] Yarmarkovich M, Warrington JM, Farrel A, Maris JM. Identification of SARS-CoV-2 vaccine epitopes predicted to induce long-term population-scale immunity. Cell Rep Med 2020;1:100036.
[233] Moulana A, Dupic T, Phillips AM, Chang J, Nieves S, Roffler AA, et al. Compensatory epistasis maintains ACE2 affinity in SARS-CoV-2 Omicron BA. 1. Nat Commun 2022;13:7011.
[234] Neverov AD, Fedonin G, Popova A, Bykova D, Bazykin G. Coordinated evolution at amino acid sites of SARS-CoV-2 spike. Elife 2023;12. e82516–e82516.
[235] Rochman ND, Faure G, Wolf YI, Freddolino PL, Zhang F, Koonin EV. Epistasis at the SARS-CoV-2 receptor-binding domain interface and the propitiously boring implications for vaccine escape. MBio 2022;13. e00135-22.
[236] Rochman ND, Wolf YI, Faure G, Mutz P, Zhang F, Koonin EV. Ongoing global and regional adaptive evolution of SARS-CoV-2. Proc Natl Acad Sci 2021:118.
[237] Zeng H-L, Dichio V, Horta ER, Thorell K, Aurell E. Global analysis of >50,000 SARS-CoV-2 genomes reveals epistasis between eight viral genes. Proc Natl Acad Sci 2020;117:31519–26.
[238] Zahradník J, Marciano S, Shemesh M, Zoler E, Harari D, Chiaravalli J, et al. SARS-CoV-2 variant prediction and antiviral drug design are enabled by RBD in vitro evolution. Nat Microbiol 2021;6:1188–98.
[239] Hie B, Zhong ED, Berger B, Bryson B. Learning the language of viral evolution and escape. Science (1979) 2021;371:284–8. https://doi.org/10.1126/science.abd7331.
[240] Wang G, Liu X, Wang K, Gao Y, Li G, Baptista-Hon DT, et al. Deep-learning-enabled protein–protein interaction analysis for prediction of SARS-CoV-2 infectivity and variant evolution. Nature 2023;29:2007–18. https://doi.org/10.1038/s41591-023-02483-5.
[241] Wong F, de la Fuente-Nunez C, Collins JJ. Leveraging artificial intelligence in the fight against infectious diseases. Science (1979) 2023;381:164–70. https://doi.org/10.1126/science.adh1114.
[242] Zvyagin M, Brace A, Hippe K, Deng Y, Zhang B, Orozco Bohorquez C, et al. GenSLMs: genome-scale language models reveal SARS-CoV-2 evolutionary dynamics. Int J High Perform Comput Appl 2023;37:683–705. https://doi.org/10.1101/2022.10.10.511571.
[243] Saucedo O, Laubmeier A, Tang T, Levy B, Asik L, Pollington T, Feldman OP. Comparative analysis of practical identifiability methods for an SEIR model[J]. AIMS Math 2024;9(9):24722–61. https://doi.org/10.3934/math.20241204.