Bulletin of Mathematical Biology (2024) 86:49
https://doi.org/10.1007/s11538-024-01275-3

Society for Mathematical Biology

ORIGINAL ARTICLE

Artificial Neural Network Prediction of COVID-19 Daily Infection Count

Ning Jiang[1] · **Charles Kolozsvary**[1] · **Yao Li**[1]

Received: 23 June 2023 / Accepted: 21 February 2024 / Published online: 1 April 2024

Abstract

This study addresses COVID-19 testing as a nonlinear sampling problem, aiming to uncover the dependence of the true infection count in the population on COVID-19 testing metrics such as testing volume and positivity rates. Employing an artificial neural network, we explore the relationship among daily confirmed case counts, testing data, population statistics, and the actual daily case count. The trained artificial neural network undergoes testing in in-sample, out-of-sample, and several hypothetical scenarios. A substantial focus of this paper lies in the estimation of the daily true case count, which serves as the output set of our training process. To achieve this, we implement a regularized backcasting technique that utilize death counts and the infection fatality ratio (IFR), as the death statistics and serological surveys (providing the IFR) as more reliable COVID-19 data sources. Addressing the impact of factors such as age distribution, vaccination, and emerging variants on the IFR time series is a pivotal aspect of our analysis. We expect our study to enhance our understanding of the genuine implications of the COVID-19 pandemic, subsequently benefiting mitigation strategies.

Keywords Covid-19 case count · Artificial neural network · Backcasting method · Infection fatality ratio

Mathematics Subject Classification 92D30 · 68T07 · 65Z05

✉ Yao Li
yaoli@math.umass.edu

Ning Jiang
ningjiang@umass.edu

Charles Kolozsvary
ckolozsvary@umass.edu

[1] Department of Mathematics and Statistics, University of Massachusetts, 710 N Pleasant St, Amherst 01003, MA, USA

1 Introduction

Since 2020, COVID-19 has infected the majority of the global population, causing nearly 7 million deaths worldwide and enormous economic losses (Organization 2023). While the severity of SARS-CoV-2 has significantly decreased due to the circulation of less virulent variants and a hybrid immunity resulting from vaccination and natural infection, COVID-19 still poses a significant threat to high-risk groups. As of late 2023, COVID-19 still causes tens of thousands hospitalizations each week in the United States alone. Over the past three years, numerous new variants with high fitness in immune escape and transmission have emerged (Harvey et al. 2021). As of now, the major threat from COVID-19 stems from the potential emergence of new and possibly more virulent variants. This underscores the importance of collecting data, improving its quality, assimilating it with models, and monitoring the circulation of SARS-CoV-2 variants. Public health agencies must stay informed about the current COVID-19 situation, including the variant composition, the number of COVID-19-related hospitalizations and deaths, as well as the percentages of people who are susceptible, recently exposed, contagious, and recently recovered from COVID-19.

Since the beginning of the COVID-19 pandemic, it has been well-known that the daily confirmed cases reported by healthcare agencies only represent a small proportion of the true daily infection count (Wu et al. 2020). Increasing testing efforts can effectively reduce the ratio of unconfirmed infection cases. The World Health Organization (WHO) recommends that the test positivity rate should be between 3% and 12%. However, determining the true infection count solely from the test positivity rate is challenging, as the distribution of people undergoing COVID-19 tests is not uniform across the population. The lack of an accurate daily infection count significantly impacts both data quality and modeling efforts, both of which are crucial to making correct predictions and mitigation strategies.

The problem of lacking high-quality data has worsened due to two main reasons. First, in 2020 and 2021, the infection fatality ratio (IFR) could be estimated by combining COVID-19-related death counts with serological surveys (Team 2022; Brazeau et al. 2022; Meyerowitz-Katz and Merone 2020). However, this method is no longer viable as nearly everyone in the world has either been infected or vaccinated, and a significant proportion of individuals have experienced multiple infections. The challenge of distinguishing between individuals who "die with COVID-19" and those who "die from COVID-19" further complicates the picture. It is evident that hybrid immunity (from infection and vaccination) and improvements in treatment have significantly reduced the IFR, but obtaining an accurate estimate has become much more difficult today. Secondly, in 2020 and 2021, most individuals who tested positive were recorded and reported by state public health agencies. However, since the spring of 2022, an increasing number of people have been conducting self-tests at home using home antigen tests, and these positive test results are no longer reported to public health agencies. Consequently, since the spring of 2022, we have become increasingly uncertain about the number of new infections occurring each day.

To address these issues, we propose the use of an artificial neural network trained with data from the period when death counts and the infection fatality ratio (IFR) were more reliable. This approach aims to predict the current state of COVID-19 circula-

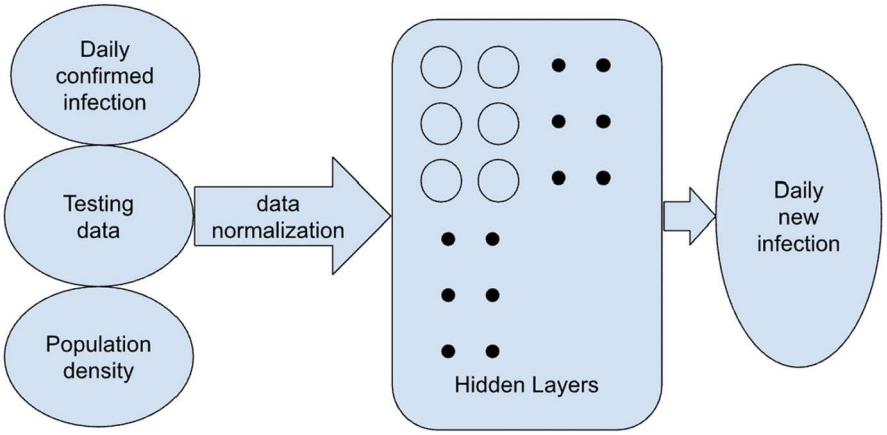

Fig. 1 Artificial neural network approximation of true daily case count. Input data includes confirmed case count, population density, and testing data. Neural network takes normalized input data (Color figure online)

tion from testing data. The COVID-19 testing can be seen as a nonlinear sampling problem, as people who go for test usually have higher risk of testing positive than the general population. However, the relation between true infection, confirmed infection, and testing volume is unknown. As depicted in Fig. 1, we enable the artificial neural network to learn the nonlinear relationship between confirmed case count, testing data, population density, and true case count. Mathematically, this is a classical function approximation problem. Once the neural network approximates this relationship, we can utilize it to predict the true COVID-19 case count when the confirmed case count and testing data are known. In addition to making predictions, the neural network can also help us understand the connection between testing data and case counts, enabling a better understanding of how many tests are necessary to limit the undercounting factor (the ratio of true cases to confirmed cases) within a certain range. We remark that although there has been numerous papers addressing unreported COVID-19 infections (Barber et al. 2022; Hortaçsu et al. 2021; Chen et al. 2022; Albani et al. 2021), estimating true infection count using machine learning approaches (Tang and Cao 2023; Guo and He 2021; Vaid et al. 2020; Dairi et al. 2021; Kamalov et al. 2022), or forecasting new infections based on existing data (Rahimi et al. 2023; He et al. 2020; Perc et al. 2020), to the best of our knowledge, the nonlinear relation between true infection count and testing data has not been properly addressed yet.

A very significant part of this paper is devoted to the estimation of the true daily case count of COVID-19, which is the most crucial step in preparing the training data. The time series of true COVID-19 cases can be estimated by various approaches such as compartment models, statistical models, time series analysis, and machine learning methods (Vaid et al. 2020; Dairi et al. 2021; Kamalov et al. 2022; Namasudra et al. 2021; Dutta et al. 2023; Chimmula and Zhang 2020; He et al. 2020; Watson et al. 2021). The quality and applicability of these studies vary a lot because much COVID-19 data has low quality and many model parameters are difficult to estimate. Many studies using compartment models also have other issues, as discussed in the

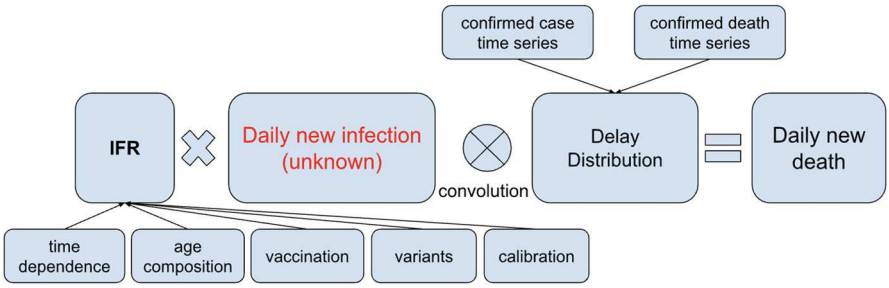

Fig. 2 A schematic diagram showing how true daily case count is estimated by solving a deconvolution problem (Color figure online)

conclusion. Generally speaking, death and hospitalization data are more trustworthy than confirmed case count. And serological survey is a more reliable way of estimating the cumulative case count. Our method of recovering the daily infection count is referred to as "backcasting" (Kevrekidis et al. 2022; Phipps et al. 2020), which relies on the death count and the infection fatality ratio (IFR). Both death count and IFR are relatively reliable in the United States because IFR can be estimated by serological surveys. We also calibrate our result with some well-recognized studies based on modeling and serological surveys.

Mathematically, the backcasting should be considered as a deconvolution problem (Miller et al. 2022; Jahja et al. 2022) rather than convolution used in some studies (Phipps et al. 2020; Sarría-Santamera et al. 2022). Since deconvolution is numerically unstable, some regularizations of high-order derivatives are necessary to accurately recover the true infection count. A schematic diagram in Fig. 2 describes our three-step approach of estimating the true daily case count. (i) The distribution of the time delay from a confirmed case to a confirmed death is recovered from the time series of daily confirmed case and daily confirmed death by solving a deconvolution problem. (ii) The IFR time series is estimated from well-recognized published data such as Team (2022). Factors such as age distribution of cases, treatment improvements, vaccination rates, and changes in variants are taken into account. (iii) We calibrate the baseline IFR for each state using the findings in well-recognized studies (Irons and Raftery 2021; Team 2022), which combine modeling and serological surveys.

The training of the neural network is inspired by the physical-informed neural network (PINN) method (Raissi et al. 2019). Since the available data only cover a relatively small region of the entire domain, the neural network exhibits limited generalization power. This limitation hampers its ability to make accurate predictions or investigate the relationship between true/confirmed cases and testing data. To overcome this challenge, we incorporate artificially generated input data and use the derivatives of the output with respect to the input data to enhance the training process. The underlying idea is that the true case count should increase with the confirmed case count and decrease with the testing volume. This concept introduces a regularization term that can be applied throughout the entire domain, as it does not rely on the output data (recovered true case count). We refer to this technique as "biology-informed regularization". Our results demonstrate that this regularization significantly improves the generalization ability of the neural network.

The paper is organized as follows. Section 2 introduces the artificial network approximation and explores the relationship between confirmed cases, testing data, and true cases in several scenarios. Section 3 explains how the artificial neural network is trained. The generation of the training set, particularly the recovery of the daily infection count, is examined in Sects. 4 and 5. The focus of Sects. 4 and 5 are backcasting method and IFR respectively. The data source of our study is introduced in Sect. 6. Finally, Sect. 7 presents the conclusion of the study.

2 Artificial Neural Network Approximation of Daily Infection Prediction

Estimating the true daily COVID-19 infection cases can be regarded as a nonlinear sampling problem. Everyday a small subset of the total population undergoes COVID-19 testing. While it may be tempting to infer the number of total infected individuals from the testing volume and the testing positivity, real-world scenarios reveal that individuals opt for COVID-19 testing due to various reasons, including experiencing symptoms or having close contact with confirmed cases. Routine tests are also conducted in many schools and workplaces. Consequently, individuals undergo COVID-19 tests have a higher likelihood of test positive than the general population. The relative risk of infection within the tested population, compared to the untested population, is nonlinearly influenced by numerous factors. As a result, traditional statistical estimation methods cannot be used to recover true infection count from testing data alone. Instead, we employ an artificial neural network to predict the true daily new infection count.

2.1 Approximation of the True Infection Count

As mentioned in the introduction, the primary goal of this paper is to utilize an artificial neural network to learn a nonlinear function:

$$I_t \approx f(I_c, \boldsymbol{\lambda}, \boldsymbol{\theta}).$$

Here, I_t represents the daily true infection count, I_c represents the daily confirmed infection count, $\boldsymbol{\lambda}$ encompasses various parameters such as testing volume, testing rate, population density, mobility, and wastewater viral RNA concentration, and $\boldsymbol{\theta}$ represents the trainable neural network parameters. By employing this function f, we can better comprehend the relationship between the undercounting factor (the ratio of daily true infections to daily confirmed infections) and other associated parameters. This understanding will aid public health agencies in obtaining insights into the current state of the COVID-19 pandemic and determining the suitable testing volume.

The parameter set $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3)$ we select in this paper consists of testing volume, testing rate per capita, and local population density. In order to facilitate the training of the neural network, certain transformations are necessary to centralize and normalize the distribution of the training data. See Sect. 3 for the full details.

The training set of our neural network includes the true daily infection count (I_t), confirmed daily infection count (I_c), testing volume (λ_1), testing rate per capita (λ_2), and population density data (λ_3) from 50 states plus Washington DC, spanning from February 29, 2020, to March 31, 2022. The input set ($I_c, \boldsymbol{\lambda}$) comes from public health data (see Sect. 6). The output set, i.e., the true infection count I_t, is estimated in Sects. 4 and 5 by using a backcasting method from daily death count.

Upon completion of the neural network training, we find a suitable parameter $\boldsymbol{\theta}^*$ that minimizes the loss function. This gives a function

$$\hat{I}_t = \hat{f}(I_c, \boldsymbol{\lambda}) := f(I_c, \boldsymbol{\lambda}, \boldsymbol{\theta}^*)$$

that describes the relationship between the undercounting factor, testing effort, and local population density. Before introducing details of the neural network training, in the following three subsections, we will demonstrate the property of the function $\hat{f}$ through in-sample validation, out-of-sample forecast, and infection prediction in some hypothetical scenarios.

2.2 In-sample Validation

To visually assess the performance of the trained neural network visually, we conduct an in-sample test using training data from October 1st, 2021, to February 1st, 2021, in six states: California, Georgia, Massachusetts, New York, Texas, and Wyoming. The comparison between predicted true cases $\hat{I}_t := \hat{f}(I_c, \boldsymbol{\lambda})$, confirmed cases I_c, and recovered true cases I_t is illustrated in Fig. 3. From the figure, the recovered true case and the predicted true case roughly follow the same trend. Both recovered true cases and predicted true cases indicate that COVID-19 tests can capture between one third and one half of the infected individuals. The United States had intensive COVID-19 testing effort during this period, with about 1.5–2 million PCR tests carried out daily. This highlights the difficulty of capturing the majority of COVID-19 cases through PCR testing.

This outcome underscores the capability of the artificial neural network to rectify inaccuracies in the training set. The recovered true cases I_t comes from backcasting the death data, which contains some randomness, holiday effects, and human error (See Sects. 4 and 5 for more detail). In particular, this period encompassed the Christmas and New Year holidays, and holiday factors may influence the accuracy of recovered cases because fewer deaths are reported during holidays. Despite our efforts to manually mitigate theses effects, the recovered true case remains imperfect. As seen in Fig. 3, occasionally the recovered true case I_t in the training set can be lower than the confirmed I_c, indicating an obvious inaccuracy. Our result shows that the artificial neural network avoids overfitting and provides a more reliable prediction of the true case count. For instance, the recovered true case in Massachusetts in January 2021 is notably low, possibly due to reporting issues in the death statistics. As a result the recovered true cases (blue line) lie below the confirmed cases (yellow line). While the artificial neural network prediction correctly shows that the true daily COVID-19 cases is around 10000. Nearly half of COVID-19 cases were undetected despite intensive testing effort. (Refer to Figs. 25 and 26 for the daily testing volume of each state.)

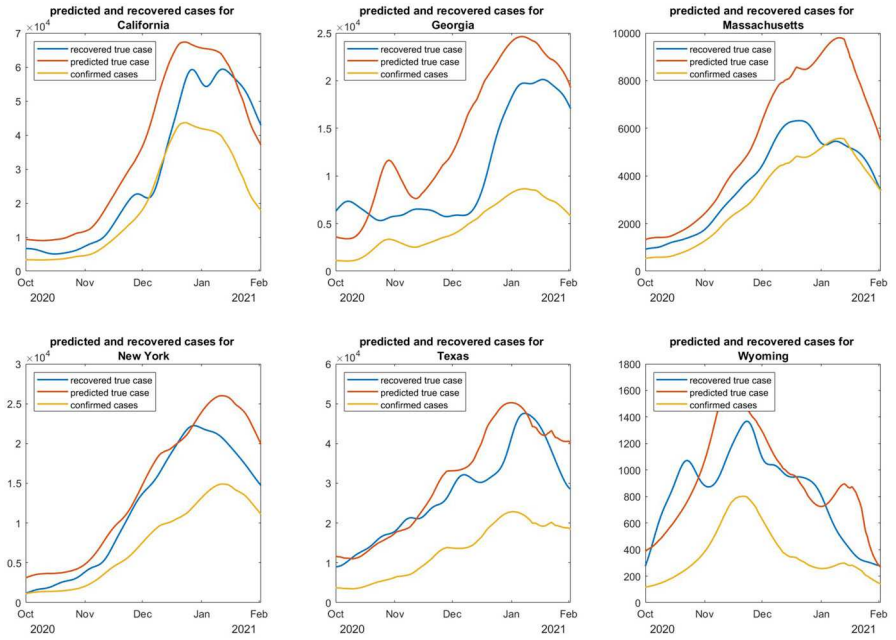

Fig. 3 Comparison of predicted true cases, recovered true cases, and confirmed cases between October 1st, 2021 and February 1st, 2021. Six panels are for California, Georgia, Massachusetts, New York, Texas, and Wyoming, respectively (Color figure online)

2.3 Out-of-Sample Forecast

To further evaluate the performance of the neural network predictions, we conducted an out-of-sample testing using the daily confirmed case and testing data from March 1st, 2022, to July 1st, 2022. This evaluation focuses on six randomly selected states California, Georgia, Massachusetts, New York, Texas, and Wyoming. These states were chosen to cover a range of parameters such as testing rate, case rate, and population density. Notably, this period witnessed a surge in the SARS-COV-2 variant Omicron BA.2 across the United States. In Fig. 4, we present a comparison between the predicted true case count and the confirmed case count during this time frame. The predicted true daily new infections greatly surpass the number of confirmed cases. Lower testing effort is generally associated with more undetected infections. The undercounting factor is notably higher in Texas and Georgia primarily due to its lower testing rate per capita in comparison to other selected states.

2.4 Exploration of Hypothetical Scenarios

One of the primary objectives of the neural network approximation is to reveal the relationship between the recovered true cases, confirmed cases, and testing effort. To explore this relationship, we establish a baseline using the testing data and daily new confirmed case data from Massachusetts, New York, and Texas on November 15th,

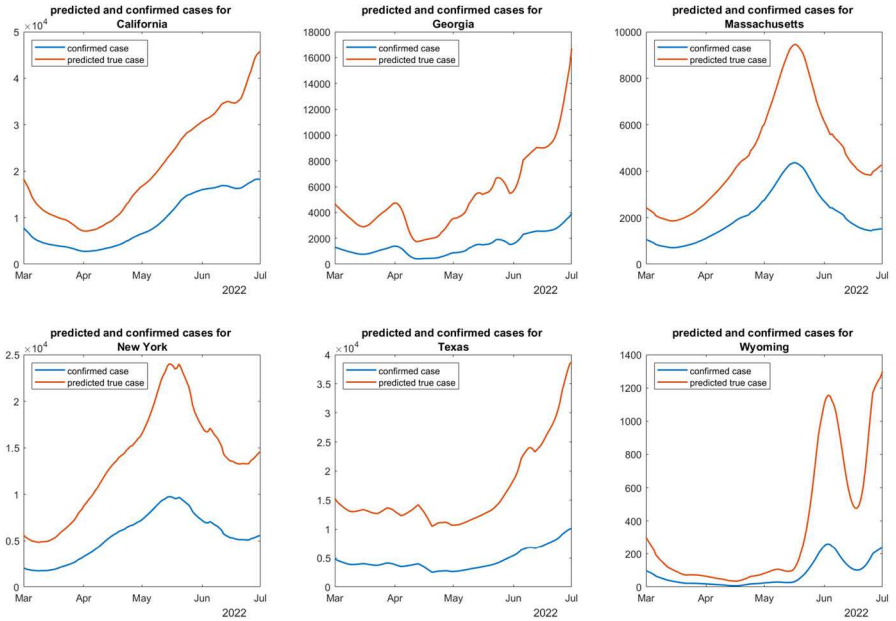

Fig. 4 Comparison of predicted true case and confirmed case between March 1st, 2022 and July 1st, 2022. Six panels are for California, Georgia, Massachusetts, New York, Texas, and Wyoming, respectively (Color figure online)

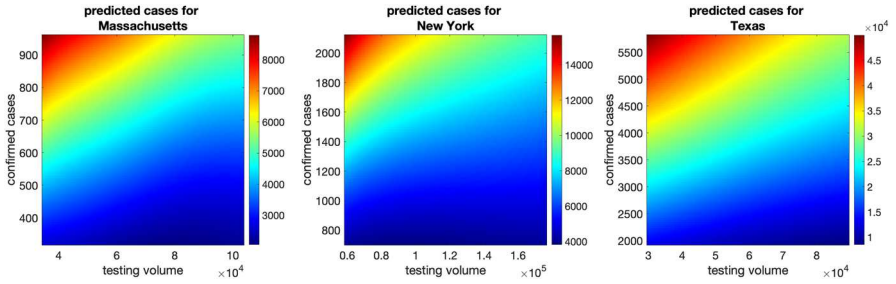

Fig. 5 Predicted case with varying confirmed case and testing volume. Three panels are for Massachusetts, New York, and Texas, respectively (Color figure online)

2020. This date is chosen because the case rate and the testing volume at that time roughly lies in the middle of the range of the dataset. Next, we investigate different hypothetical scenarios by modifying the testing volume by a factor of x_1 from the baseline and changing the confirmed cases by a factor of x_2 from the baseline. In other words, we plot $\hat{f}(x_1 I_c, x_2\lambda_1, x_2\lambda_2, \lambda_3)$ for the trained neural network approximation function $\hat{f}$. The values of x_1 and x_2 are selected from a 100×100 equi-spaced 2D grid within the range of $[0.5, 1.5] \times [0.5, 1.5]$. We generate a total of 10^4 new scenarios, which are then input into the trained artificial neural network predictor $\hat{f}$. The results are illustrated in Fig. 5. As anticipated, the predicted true cases increase with the confirmed case count and decrease with the testing volume.

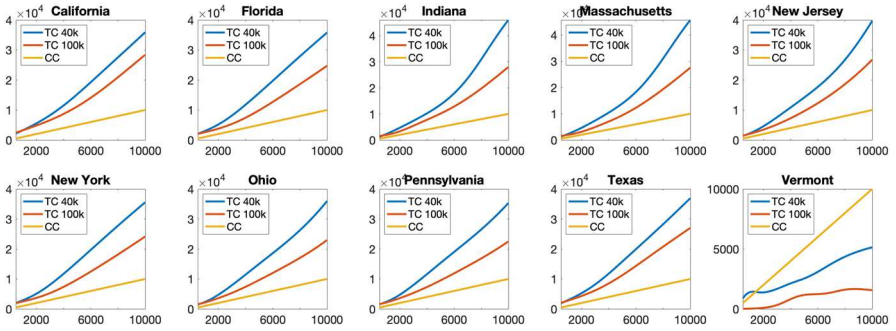

Fig. 6 Predicted case with varying confirmed case and two fixed testing volume scenarios. TC 40k and TC 100k mean recovered true case with 40,000 and 100,000 daily tests respectively. CC means confirmed true case (Color figure online)

To gain a clearer understanding of the relationship between the recovered true cases, confirmed cases, and testing efforts, we conduct an analysis considering two distinct scenarios. In scenario A, 40,000 tests are performed, while in scenario B, 100,000 tests are conducted within a state. The number of positive cases I_c (i.e., confirmed cases) is varied from 500 to 10,000. We examine the predicted true case count $\hat{I}_t = \hat{f}(I_c, \boldsymbol{\lambda})$ in ten different states: California, Florida, Indiana, Massachusetts, New Jersey, New York, Ohio, Pennsylvania, Texas, and Vermont. The results are depicted in Fig. 6.

The findings reveal that, for the first nine states, under the same testing effort, the recovered case count exhibits a super-linear growth pattern in relation to the confirmed case count. For instance, if 10,000 positive cases are detected out of 40,000 tests, the recovered true case count is approximately four times higher than the confirmed case count. As the testing effort increases and 10,000 positive cases are identified out of 100,000 tests, the recovered true case count is only about 2–3 times higher than the confirmed case count. The last panel of Fig. 6 demonstrates a scenario where the trained neural network fails to accurately predict the true recovered cases. This discrepancy arises because the highest recorded daily test count in Vermont is only 12,000. As a result, the two scenarios tested here are significantly different from the training set that the neural network has learned from.

Another interesting observation is that even with sufficient testing, the recovered true case count does not approach zero when the confirmed case count reaches zero. This can be attributed to a combination of factors, including the nature of COVID-19 testing and potential data artifacts. On one hand, regardless of the number of tests conducted by a state, cases in certain underserved communities and remote areas may remain undetected. Consequently, this leads to a non-zero extrapolation of the predicted true case count. On the other hand, there is a possibility of over-counting deaths that are not actually caused by COVID-19 as COVID-19 deaths. Although the impact of these over-counted deaths on the overall pandemic is relatively small, they can contribute a non-negligible proportion to the reported death count during periods when the case count is very low (such as in spring 2021 and spring 2022). Since the true case count is derived from the daily death count in our analysis, this factor may also inflate the estimated daily case count when the daily confirmed case count is low.

3 Artificial Neural Network Training

3.1 Training Data Preparation

As previously mentioned, the objective of the neural network approximation is to find a suitable parameter $\boldsymbol{\theta}$ to fit the function

$$I_t = f(I_c, \boldsymbol{\lambda}, \boldsymbol{\theta})$$

to the training data, where I_t and I_c represent the normalized recovered true case rate and confirmed case rate, respectively. The parameters $\boldsymbol{\theta}$ correspond to the neural network parameters, while $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3)$ represents the input data normalized from testing volume, testing rate, and population density. Both daily confirmed case count I_c and the testing data comes from the Coronavirus Resource Center of Johns Hopkins University (Center 2023). The true infection count is recovered by backcasting from the death count with a careful estimation of the IFR considering multiple factors. See Sects. 4 and 5 for details.

To ensure proper normalization of the data, we apply the following nonlinear transformations to the input and output data. Let Pop denote the local population (state population). Recall that I_t, I_c, λ_1, λ_2, and λ_3 are recovered daily true case count, daily confirmed case count, daily test volume, daily test rate per capita, and the local population density, respectively. Note that $\text{Pop} = \lambda_2/\lambda_3$ is a function of the input parameters. We utilize the following transformations to normalize data:

$$\begin{aligned}
\tilde{I}_c &= 50\sqrt{I_c/\text{Pop}} \\
\tilde{I}_t &= 25\sqrt{I_t/\text{Pop}} \\
\tilde{\lambda}_1 &= 0.05\sqrt[3]{\lambda_1} \\
\tilde{\lambda}_2 &= 200\lambda_1/\text{Pop} \\
\tilde{\lambda}_3 &= 0.2\log\lambda_3\,.
\end{aligned}$$

In other words, the artificial neural network, denoted by $\text{NN}_{\boldsymbol{\theta}} : \mathbb{R}^4 \to \mathbb{R}$, takes the value $(\tilde{I}_c, \tilde{\lambda}_1, \tilde{\lambda}_2, \tilde{\lambda}_3) \in \mathbb{R}^4$ and maps to $\tilde{I}_t \in \mathbb{R}$. The full form of the approximation function $f(I_c, \boldsymbol{\lambda}, \boldsymbol{\theta})$ is

$$I_t = \frac{\text{Pop}}{625}\text{NN}_{\boldsymbol{\theta}}(\tilde{I}_c, \tilde{\lambda}_1, \tilde{\lambda}_2, \tilde{\lambda}_3) := f(I_c, \boldsymbol{\lambda}, \boldsymbol{\theta}).$$

The function f is well defined because variables Pop, $\tilde{I}_c, \tilde{\lambda}_1, \tilde{\lambda}_2, \tilde{\lambda}_3$ are all functions of I_c and $\boldsymbol{\lambda}$.

This results in a training data set $\{(X_i, y_i)\}_{i=1}^N$, where $X_i = (\tilde{I}_c(i), \tilde{\lambda}_1(i), \tilde{\lambda}_2(i), \tilde{\lambda}_3(i))$ and $y_i = \tilde{I}_t(i)$. The nonlinear transformation shifts the distribution of training data to a suitable range for the artificial neural network. See Fig. 27 in the Appendix for the distribution of all components of the training data.

3.2 Neural Network Architecture and Training

In this paper, we employ a feed-forward neural network $\mathrm{NN}_{\boldsymbol{\theta}}$ composed of six layers to approximate the true case count I_t. The network architecture consists of layers with widths of 64, 128, 128, 128, 64, and 16, respectively. Each layer incorporates a sigmoid activation function and includes an L^2 regularization term with a magnitude of 0.005. The input layer takes the input $X_i = (\tilde{I}_c(i), \tilde{\lambda}_1(i), \tilde{\lambda}_2(i), \tilde{\lambda}_3(i))$, while the output layer approximates $y_i = \tilde{I}_t(i)$.

By the universal approximation theorem, any continuous function can be well-approximated by a sufficiently large fully connected feed-forward artificial neural network (Kidger and Lyons 2020; Maiorov and Pinkus 1999). Here we choose this network architecture based on its successful application in solving high dimensional partial differential equations in our prior studies (Zhai et al. 2022). Therefore, we know that this neural network is large enough to well approximate a continuous function in $\mathbb{R}^4$. Through experimental tests, we find that the verification loss remains at 0.03 even when adding an additional layer to the network. Given the relatively small size of our data, a larger architecture is unnecessary. We choose the sigmoid activation function instead of ReLU because the ReLU activation function exhibits discontinuities in its first-order derivative. Since we require the first-order derivative for regularization during training, the sigmoid activation function proves to be a suitable choice. The neural network training uses the "alternating Adam" approach described below. The loss is stabilized at around 0.03 after 25 training epochs. The neural network training is carried out on a MacBook Pro laptop. The training time is less than four minutes.

The loss function used in this study comprises two components: the classical mean squared error (L_1) and a penalty term (L_2) for regularization. The L_1 term quantifies the mean squared difference between the neural network prediction $\mathrm{NN}_{\boldsymbol{\theta}}(X_i)$ and the transformed observed daily true case count y_i:

$$L_1 = \frac{1}{N} \sum_{i=1}^{N} \left(\mathrm{NN}_{\boldsymbol{\theta}}\left(X_i\right) - y_i\right)^2 .$$

However, due to the limited coverage of the observed data in the 4D domain of $\mathrm{NN}_{\boldsymbol{\theta}}$, training solely on L_1 is insufficient to effectively capture the relationship between true cases, confirmed cases, testing volume, and testing rate. For instance, when varying the confirmed cases and testing data from a specific day, there is a noticeable discrepancy in the neural network's predictions. This is illustrated in Fig. 7, where the predicted cases do not exhibit a monotonic increase with confirmed cases or a decrease with testing volume. Moreover, the training results lack robustness, as demonstrated in Fig. 7 with two different training results yield distinct true case count profiles for Massachusetts and New York.

Therefore, to address these issues, we incorporate the concept of a Physics-informed neural network (Raissi et al. 2019) and introduce a *biology-informed regularization* based on three principles.

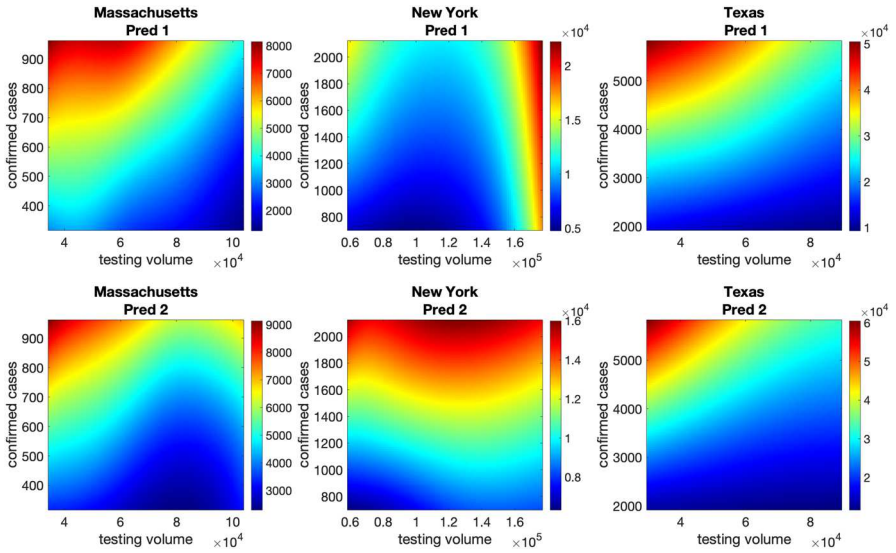

Fig. 7 Predicted case with varying confirmed case and testing volume without using regularization. Input data is identical to that of Fig. 5. Two rows are from the prediction of two different training results (Color figure online)

Firstly, we enforce that the true case count must be greater than the confirmed case count. This is achieved by including a term that penalizes the deviation between the predicted true case count and the confirmed case count.

Secondly, we ensure that the partial derivative of the true case count with respect to the confirmed case count is positive, as a higher confirmed case count implies a higher true case count. We incorporate this constraint by penalizing the negative partial derivatives.

Lastly, we aim to capture the negative relationship between the true case count and the testing rate. This is because an increased testing rate, given a fixed number of confirmed cases, suggests a lower number of true infections. To enforce this relationship, we introduce a term that penalizes positive partial derivatives of the true case count with respect to the testing rate.

By incorporating these penalty terms, the overall regularization term L_2 becomes

$$L_2 = \frac{1}{N}\sum_{i=1}^{N}[l_1(i) + l_2(i) + l_3(i)]$$

where $l_1(i), l_2(i)$, and $l_3(i)$ are the three penalty terms defined as:

$$l_1(i) = 5\max\left(\tilde{I}_c(i) - 2\mathrm{NN}_{\boldsymbol{\theta}}(X_i), 0\right)$$
$$l_2(i) = 8\max\left(-\frac{\partial \mathrm{NN}_{\boldsymbol{\theta}}}{\partial \tilde{I}_c}(X_i), 0\right)$$

$$l_3(i) = 12 \max\left(\frac{\partial \mathrm{NN}_{\boldsymbol{\theta}}}{\partial \tilde{\lambda}_1}(X_i), 0\right).$$

To compute the partial derivative terms in L_2, we can utilize backpropagation, a built-in function of TensorFlow (`tf.gradient`). As the training set derived from real-world data may not sufficiently cover the entire input domain, and since the loss term L_2 is independent of the output, we address this issue by uniformly sampling an additional set of $M = 50,000$ points within the 4D box that contains the original training set $\{(X_i, y_i)\}_{i=1}^N$. These new points, denoted as $\{Z_j\}_{j=1}^M$, provide a more representative distribution as they uniformly cover the 4D domain. Training the penalty term L_2 using the training set $\{Z_j\}_{j=1}^M$ will help the neural network "learn" and properly capture the three constraints described above.

To address the differences in scales between L_1 and L_2 and the fact that they are trained on different sets, a training strategy inspired by the "Alternating Adam" approach proposed in Zhai et al. (2022) is employed. During each training step, a random batch is sampled from the original training set $\{(X_i, y_i)\}_{i=1}^N$, and the Adam optimizer is applied to optimize L_1 with respect to this batch. Then, another random batch is sampled from the training set $\{Z_j\}_{j=1}^M$, and the Adam optimizer is used to optimize L_2 with respect to this batch. The learning rates for the Adam optimizers are set to 0.0003 for L_1 and 0.0005 for L_2. This approach effectively trains both loss functions simultaneously, regardless of their scales. The inherent scaling-invariance of the Adam optimizer ensures that the optimization is performed properly for each loss function (Kingma and Ba 2014).

To monitor the training progress and prevent overfitting, a testing set is randomly selected, consisting of 20% of the original training set $\{(X_i, y_i)\}_{i=1}^N$. The losses on both the training set and the testing set are observed during training. The training process stops after 25 epochs to avoid overtraining. Each epoch involves training both loss functions using 458 batches with a batch size of 32.

4 Generation of Training Set I: Backcasting from Daily Death Count

A well-constructed training set plays a crucial role in the success of a neural network prediction project. As discussed before, the input data consisting of the daily confirmed cases, testing volume, testing rate, and population density is relatively straightforward, as this information can be sourced from public health agencies or online databases, such as the database from the Coronavirus Resource Center of Johns Hopkins University (Center 2023). However, the true daily new cases I_t, which serve as the output of the artificial neural network, are unknown. Therefore, a significant portion of our efforts in generating the training set is focused on recovering the true daily new cases, a topic that will be extensively discussed in the next two sections.

The basic concept behind recovering the daily new cases is that

$$\text{daily new case } * \text{delay} \times \text{IFR} = \text{daily new death},$$

where IFR represents the infection fatality ratio (IFR) and $*$ denotes convolution. Hence, our strategy involves first utilizing the daily new confirmed case count and the daily new death count to infer the distribution of the delay from a confirmed COVID-19 case to a confirmed COVID-19 death. Subsequently, we utilize information regarding age composition and vaccination rates to estimate the time series of the IFR.

4.1 Data Processing

It is widely recognized that COVID-19 data reporting is subject to various factors, including weekend effects, holiday effects, noise, and potential human errors in reporting. As a result, the initial step in our analysis involves preprocessing the daily confirmed cases and reported deaths. This data processing procedure comprises the following five steps:

1. Correcting human errors. In many states, the COVID-19 data is affected by artificial data backlogs, where the case count and/or death count of multiple days are reported on a single day. Figures 17 and 18 demonstrate raw case and death count. It is easy to see that human error causes multiple artificial spikes. Sometimes thousands backlogged death accumulated in more than a year are reported in a single day, such as the top left corner of Fig. 20 (Missouri). Through extensive testing, we identified irregular reporting patterns by flagging days where the reported case/death count is at least twice the average of the previous 8 days. The excess cases/deaths resulting from backlogs are then uniformly redistributed over the previous L days. The value of L is chosen to ensure that the redistribution for each day does not exceed 65% of the average daily cases/deaths.
2. Removing weekend factor. Most states report less cases and deaths in the weekends, which causes significant periodicity of the data. This problem can be solved by taking the 7-day average.
3. Removing holiday factor. COVID-19 data reporting during Thanksgiving and Christmas/New Year periods is highly irregular due to reporting delays during these holidays. To address this issue, we employ a linear function to bridge the data before and after a specific time window. The discrepancy between the actual reported data and the linear function is then offset by redistributing the corresponding cases/deaths from the day immediately following this time window. Since the training set includes only four holidays, all time windows are manually adjusted to effectively mitigate the impact of the holiday factor.
4. Smoothing data. We use the LOESS regression method (Cleveland and Devlin 1988) to smooth the data. The smoothing window extends from 7 days before to 28 days after each data point.
5. Addressing negative fluctuations: Following the LOESS smoothing process, there is a possibility of the initial phase of case/death fluctuating below zero. To resolve this issue, we utilize an exponential function to fit the 7-day average data of the first L days (starting on January 22, 2020) for each state. Here, L represents the date of the peak of the first wave during the spring of 2020. The initial case/death counts are then replaced with the values obtained from this exponential fit.

The daily confirmed case and daily death data after processing of each state are demonstrated in Figs. 19 and 20. It is easy to see that the data become significantly smoother and more readable than the raw data in Figs. 17 and 18.

4.2 Deconvolution and Regularization

The daily reported deaths attributed to Covid-19 can be viewed as a *convolution* of the time series of fatal infections with a delay distribution. Specifically, the delay time from a confirmed case to a reported death, denoted as Δ, follows an unknown distribution:

$$\mathbb{P}[\Delta = i] = \delta_i.$$

The number of confirmed deaths in the United States on day n, denoted as D_n, can be expressed as

$$D_n = \sum_{i=0}^{n} I_i \delta_{n-i} \xi,$$

where I_i represents the confirmed cases in the United States on day i, and ξ represents the case fatality rate (CFR). Thus, the process of recovering fatal infections from reported deaths is a *deconvolution* operation. (Note that a deconvolution is different from a convolution in the other direction as Phipps et al. (2020) does, which tends to overly smooth the time series of the true infection count.) Based on the findings presented in Flaxman et al. (2020), Phipps et al. (2020), we assume that the delay distribution, $\{\delta_i\}$, follows a gamma distribution characterized by two unknown parameters, α and β.

Let N denote the duration of available data. The convolution problem can be expressed in matrix form as

$$P_N(\alpha, \beta)\vec{I}\xi = \vec{D}, \tag{1}$$

where $\vec{I}$ is a column vector of length N containing the number of confirmed infections each day, $\vec{D}$ is a similar column vector containing the number of reported deaths, and $P_N(\alpha, \beta)$ is an $N \times N$ square matrix that represents the conditional probability of death on each day. Specifically, the entry in column i and row j represents the probability that a newly confirmed COVID-19 patient on day i eventually dies on day j, conditioned on the assumption that this infection is fatal. We follow Phipps et al. (2020) by assuming that the time duration between confirmed case and confirmed death follows a Gamma distribution with unknown parameters α and β. To simplify the computation, we make the assumption that the maximum delay is 35 days. Consequently, each column of $P_N(\alpha, \beta)$ contains at most 35 non-zero entries (ranging from θ_0 to θ_m, where $m = 34$).

In other words, we have

$$P_N(\alpha, \beta)_{ij} = \begin{cases} \delta_{i-j} & \text{if} \quad 0 \le i - j \le 34 \\ 0 & otherwise \end{cases}$$

where

$$\delta_i = \mathbb{P}[i \le Z \le i + 1]$$

for a Gamma distributed random variable Z with parameters α and β.

One might attempt to find suitable parameters by minimizing $\| P_N(\alpha, \beta)^{-1} \vec{D}\xi - \vec{I} \|_2^2$ over all possible values of α and β. However, this approach is not feasible for two reasons. Firstly, the deconvolution process is known to be unstable due to the ill-conditioned nature of $P_N(\alpha, \beta)$ (Miller et al. 2022; Jahja et al. 2022). Even small noise in $\vec{D}$ can lead to significant amplification during matrix inversion. Secondly, the case fatality ratio ξ is unknown. Thus, *regularization* is necessary to prevent excessive fluctuations in the recovered $\vec{I}$. The optimization problem also involves the recovery of the unknown ξ.

The regularization is achieved by incorporating penalty terms for the second and fourth order derivatives. Two matrices, namely R_2 and R_4, are employed to discourage excessive fluctuations in the time series of confirmed cases. Each row of R_2 is responsible for regularizing one entry of $\vec{I}$ (except the first and last ones). More precisely, R_2 has the form

$$R_2 = \lambda_2 \begin{bmatrix} 1 & -2 & 1 & & & \\ & 1 & -2 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & -2 & 1 \end{bmatrix} \in \mathbb{R}^{(N-2)\times N}.$$

Similarly, matrix R_4 regularizes the fourth order derivative of entries of $\vec{I}$ (except the first two entries and the last two entries). It has the form

$$R_4 = \lambda_4 \begin{bmatrix} 1 & -4 & 6 & -4 & -1 & & & \\ & 1 & -4 & 6 & -4 & 1 & & \\ & & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & & 1 & -4 & 6 & -4 & -1 \end{bmatrix} \in \mathbb{R}^{(N-4)\times N}.$$

With the regularization and the tuning parameter λ, we can utilize the modified delay matrix P to perform a reliable deconvolution of the reported death time-series into the corresponding fatal infections. This allows us to estimate the daily confirmed infection count by solving the following system using the least squares method:

$$\begin{bmatrix} P_N(\alpha, \beta)\xi \\ \lambda_2 R_2 \\ \lambda_4 R_4 \end{bmatrix} \vec{I} = \vec{D}. \tag{2}$$

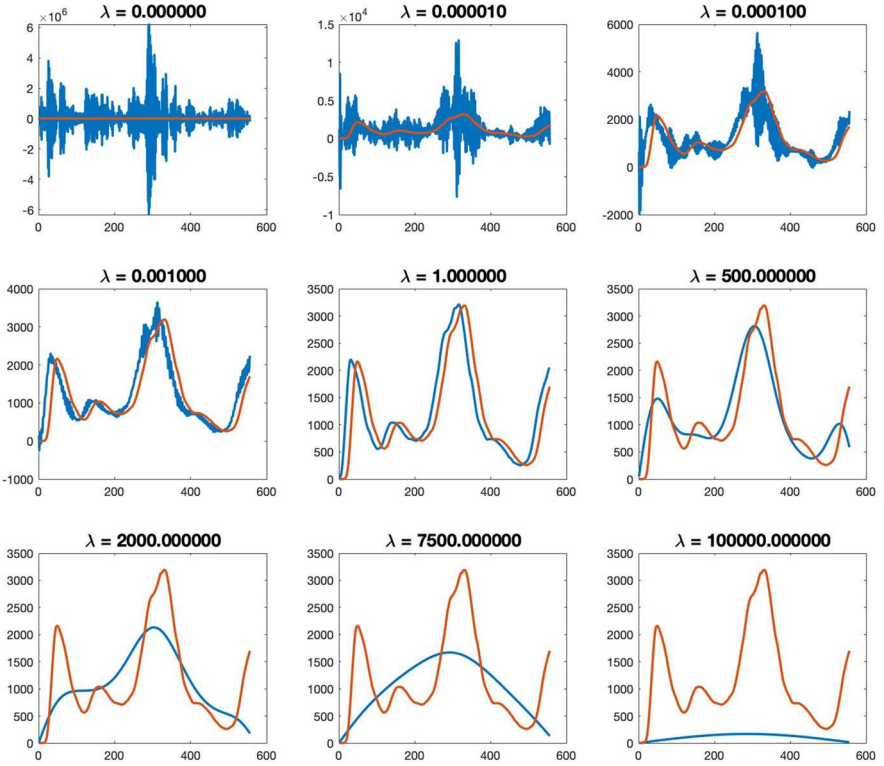

Fig. 8 The red curve shows the reported death time series in the United States between 03/01/20 and 11/30/21, and the blue curve is the deconvolve time series of fatal infections when some arbitrary matrix P is constructed with values of $\alpha = 30$ and $\beta = 0.5$. For the sake of simplification, $\lambda_4 = 0$ in all plots. $\lambda = \lambda_2$ changes from 0 (no regularization) to 10^5 (too much regularization) in six plots (Color figure online)

Figure 8 demonstrates the effectiveness of regularization and how the variation of λ_2 and λ_4 impacts the recovered least squares solution of equation (2).

Since the parameter ξ (CFR) is also unknown, it must be solved in the optimization problem together with α and β. After conducting several tests, we determine that $\lambda_2 = 0.5$ and $\lambda_4 = 2$ are suitable coefficients for the regularization matrices. Additionally, we only focus on minimizing the difference between the observed $\vec{I}$ and the least squares solution after a period of 120 days from the start date (March 1st, 2020). This choice is made because the testing was limited during the initial few months of the pandemic, and widespread testing became available in the summer of 2020, stabilizing the case fatality rate. This gives the optimization problem

$$\min_{\alpha,\beta,\xi} \|\mathcal{P}_{120}(\vec{I} - \hat{I}(\alpha, \beta, \xi))\|_2^2 \tag{3}$$

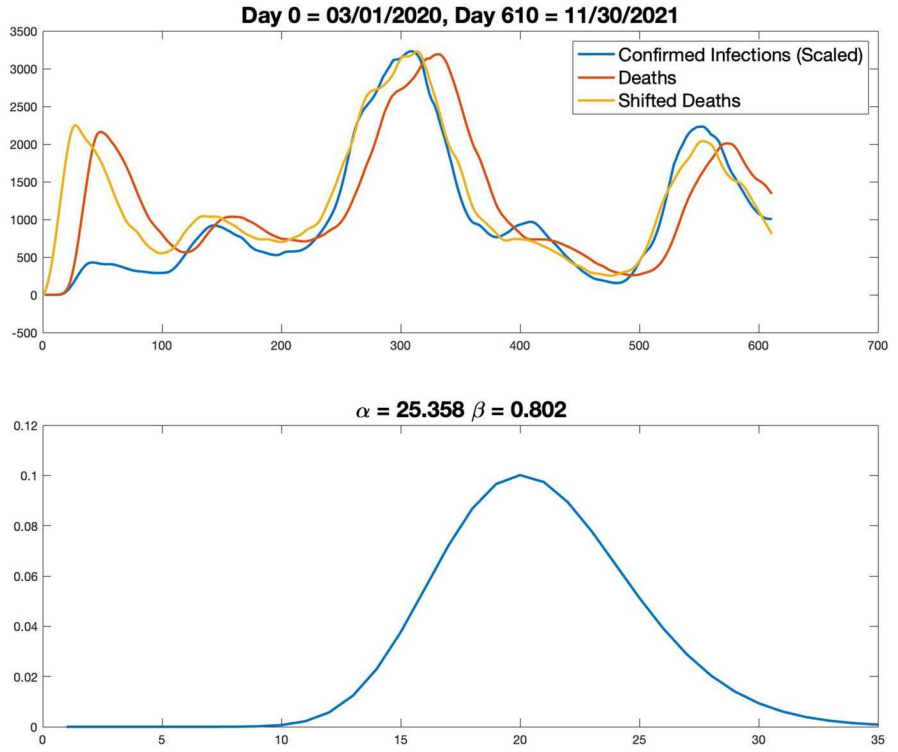

Fig. 9 The result of the deconvolution is the yellow curve (found using $\lambda_2 = 0.5$ and $\lambda_4 = 2$). The delay distribution is shown in the lower panel (Color figure online)

where $\hat{I}$ is the least square solution of

$$\begin{bmatrix} P_N(\alpha, \beta)\xi \\ \lambda_2 R_2 \\ \lambda_4 R_4 \end{bmatrix} \hat{I} = \vec{D},$$

the projection matrix $\mathcal{P}_{120}$ cuts off the first 120 entries of the vector. We implemented the optimization procedure using the `fmincon` function in MATLAB, and the results are displayed in Fig. 9. The daily confirmed case count and daily death count of the United States come from the Coronavirus Resource Center of Johns Hopkins University (Center 2023). As illustrated in the figure, the recovered case count from the death count shows a reasonable match with the confirmed case count, particularly after the availability of widespread testing in the summer of 2020. The optimal values obtained for α and β are $\hat{\alpha} = 25.358$ and $\hat{\beta} = 0.802$, respectively. These values indicate that the expected value of the delay between a fatal infection and the reported death is approximately $\hat{\alpha}\hat{\beta} = 20.337$ days.

One important observation from this deconvolution process is that the CFR ξ does not show significant change from mid-2020 when testing became widely available to late 2021 before the surge caused by Omicron variant. As seen in Fig. 9, the yellow

line and blue line remains parallel during this time period. Therefore, we only address the change of infection fatality rate for the Omicron variant. See Sect. 5 for details

4.3 Recovery of Daily Infection Count

After obtaining the optimal values $\hat{\alpha}$ and $\hat{\beta}$, we can recover the true daily new infection count by solving the following least squares problem:

$$\begin{bmatrix} P_N(\alpha, \beta)\Gamma_{ifr} \\ c\lambda_2 R_2 \\ c\lambda_4 R_4 \end{bmatrix} \hat{I} = \vec{D}, \tag{4}$$

where Γ_{ifr} is a diagonal matrix whose entries represent the time series of the infection fatality rate (IFR), and c is the average value of the IFR. Note that the IFR is typically a small value around 0.01. Without multiplying c, we would overly regularize the deconvolution problem. The solution to this least squares problem, denoted by $\tilde{I}$, represents the time series of the recovered infection count.

As seen from the problem (4), the next crucial data required is the time series of the infection fatality rate (IFR). The IFR is influenced by various factors, including the overall healthiness of the population, treatment methods, age composition of cases, vaccination rate, and the presence of variants. In the following section, we will discuss these factors in detail.

5 Generation of Training Set II: Estimation of Infection Fatality Ratio (IFR)

The infection fatality ratio (IFR) of a state at time t can be expressed as

$$\text{IFR}(t) = \text{IFR}_b \times \text{IFR}_R(t),$$

where IFR_b is the baseline IFR and $\text{IFR}_R(t)$ is the relative change in IFR over time. The baseline IFR is calibrated using well-acknowledged data, as discussed in the previous subsection. The time series $\text{IFR}_R(t)$ represents the relative changes in the IFR due to various factors, including improvements in treatment, changes in the age composition of cases, vaccination efforts, and the emergence of different variants. More precisely, $\text{IFR}_R(t)$ is represented by

$$\text{IFR}_R(t) = \text{IFR}_T(t) \times \text{IFR}_A(t) \times \text{IFR}_V(t) \times \text{IFR}_O(t),$$

where $\text{IFR}_T(t)$ is the relative reduction of IFR due to the improvement of treatment, IFR_A is the relative change in IFR due to the age composition of cases, $\text{IFR}_V(t)$ is the relative reduction in IFR due to vaccination, and $\text{IFR}_O(t)$ is the reduction in IFR due to the Omicron variant. At the baseline (July 1st, 2020), all four factors of the IFR are assumed to be equal to 1.

We remark that the real world IFR depends on many other factors that are not addressed in our paper, such as variants in addition to Omicron, protection from prior infections, seasonality, and the capacity of healthcare systems. In order to proceed, we also need to make many assumptions such as the linear extrapolation of IFR baseline and the mortality rate of population with advanced age. We also assume that the time series of IFR of each state is roughly proportional to that of the US average. While we make every effort to estimate all quantities accurately using the data that is available to us, any real-world estimation inherently comes with numerous hypotheses and limitations.

On the other hand, as seen in Fig. 9 (the yellow and blue plots), the ratio of confirmed death to confirmed infection remains largely stable with a very slow decrease from mid-2020 (when the testing became widely available) all the way until late-2021 (when the Omicron variant became dominant). Noting that the majority of infected people came from unvaccinated and uninfected group before the surge of Omicron variant. We believe this evidence supports our assumptions that the IFR baseline can be linearly extrapolated (Sect. 5.1) and that the variants before Omicron did not significantly change IFR in the real world (Sect. 5.4).

5.1 Time Dependence of IFR Baseline

Since the beginning of the pandemic, significant advancements have been made in the treatment of COVID-19. The relative reduction in the infection fatality rate (IFR_T) due to treatment improvements is obtained from the study (Team 2022). We gather IFR estimates for the United States on April 15, 2020, July 15, 2020, October 15, 2020, and January 1, 2021. To estimate the values between April 15, 2020, and January 1, 2021, cubic interpolation is employed. Linear extrapolation is used for estimating IFR_T before April 15, 2020, and after January 1, 2021. The extrapolation is halted in March 2022 when oral antiviral treatments become widely available. Linear interpolation is no longer suitable after this time. At the end of our estimation, the final $\text{IFR}_T(t)$ is approximately 0.005 before rescaling to the baseline. This estimate may be slightly conservative as monoclonal antibody treatments became widely available in 2021. However, it is challenging to estimate the reduction in IFR for each category of treatment method. The calibration performed in the previous subsection, utilizing serological survey data, partially addresses this issue. The plot of $\text{IFR}_T(t)$ is shown in Fig. 10 (left).

5.2 Change of Age Compositions

Unlike many other pathogens, age is the most significant risk factor for COVID-19. The disease poses a considerable risk to individuals in advanced age groups. As illustrated in Fig. 10 (right), based on data from Team (2022), the infection fatality rate (IFR) for those aged 85 and above is thousands of times higher compared to younger age groups. Moreover, due to changes in public health policies and events such as nursing home outbreaks and school reopenings, the age composition of confirmed COVID-19 cases

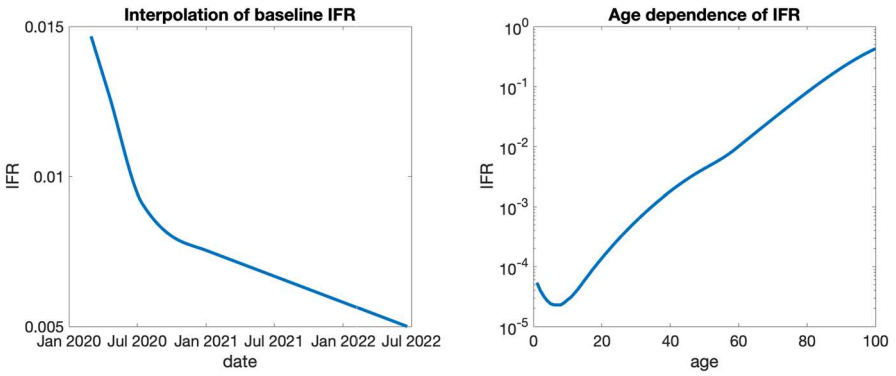

Fig. 10 Left: time dependent baseline IFR. Right: IFR for different age group in log-linear plot (Color figure online)

has varied significantly throughout the pandemic. Therefore, it is crucial to estimate the age composition of COVID-19 cases for each state.

Case rates for seven different age groups (under 20, 20–29, 30–39, 40–49, 50–64, 65–74, and 75+) in ten different Health and Human Services (HHS) regions are obtained from the CDC COVID-19 patient database (Disease Control 2023a). For the under 20 age group, which is further divided into many groups in the CDC patient database, we use the case rate of ages 12–15 to represent the entire under 20 age group. This approximation has a minimal impact on the overall population IFR since the IFR for the under 20 age group is very low. The case rates at the beginning of each month are collected and interpolated using the modified Akima algorithm (Akima 1970, 1974). The advantage of the modified Akima algorithm is its ability to minimize overshooting or undershooting when data changes dramatically. However, during the Omicron surge at the end of 2021, weekly data is utilized as the case rates exhibit significant fluctuations during this period. The interpolated case rates for all age groups and regions are presented in Fig. 21 in the Appendix.

The time series $\mathrm{IFR}_A(t)$ can be obtained by calculating a weighted average of the case rate and the age group IFR. The age group IFR is derived from a weighted average of the age-specific IFR values reported in the study (Team 2022) and the population of each 5-year age group based on the 2020 US Census data. However, due to the significantly higher IFR in the 85+ age group, it needs to be estimated separately with a better approximation of number of individuals at each age. According to Figure 3 of Easton and Hirsch (2008), for people with age ≥ 80, the mortality rate has a linear growth. Therefore, in this estimation, we assume that the number of individuals at each age 85 or older decreases as the exponential of a quadratic function with unknown parameters. As there are approximately 97,000 individuals in the United States aged over 100 years, unknown parameters of the quadratic function can be fitted. The IFR of the 85+ age group is obtained through a weighted average of the IFR values for each specific age provided in Team (2022) and the estimated number of individuals in each age group based on the fitting.

5.3 Change of Vaccination Rate

To assess the relative change in IFR due to vaccination, represented by the time series IFR_V, we need to estimate the relative risk of cases and deaths for the vaccinated group compared to the unvaccinated group. This can be achieved by analyzing CDC data provided by Disease Control (2023b), which provides case rates and death rates for each age group among both the vaccinated and unvaccinated populations.

The CDC vaccination data covers five age groups: 18–29, 30–49, 50–64, 65–79, and 80+. Vaccination for individuals under 18 years old does not significantly impact the overall IFR due to their relatively low risk. It's important to note that the CDC data is reported on a weekly basis. To match the daily basis of our analysis, we use cubic spline interpolation to convert the data to daily values.

For a specific age group i, we can determine the relative risk of cases and deaths by comparing the case and death rates of the unvaccinated group to those of the vaccinated group. Let $R_c(i)$ and $R_d(i)$ denote the relative risk of cases and deaths, respectively, for age group i. Additionally, let $\alpha(i)$ represents the vaccination rate of age group i, and d_i denotes the death rate of this age group. We can then determine the proportion of deaths contributed by the vaccinated group, denoted by $\beta(i)$,

$$\beta(i) := \frac{\alpha(i)}{\alpha(i) + (1 - \alpha(i))R_d(i)}.$$

Considering the IFR of age group i as $\text{IFR}(i)$, we can then calculate the relative reduction of IFR due to vaccination using the expression

$$\frac{\sum_i \frac{d_i}{\text{IFR}(i)}}{\sum_i \frac{d_i \beta(i) R_d(i)}{R_c(i)\text{IFR}(i)} + \sum_i \frac{d_i(1-\beta(i))}{\text{IFR}(i)}}.$$

Performing this calculation for each day allows us to obtain the time series IFR_V. The daily vaccination rates for each age group can be obtained from CDC data set (Disease Control 2023b).

5.4 Change of Variants

The final step is to account for the impact of variants on the infection fatality rate (IFR). As discussed in the beginning of this section, based on the observations shown in Fig. 9, neither the Alpha variant nor the Delta variant significantly altered the case fatality ratio in the United States, despite some studies indicating that the Delta variant may be more intrinsically virulent. This may be because high risk groups are already vaccinated when the Delta variant surged in United States. Significantly lower risk of vaccinated and infected group after contracting the Delta variant may also contribute.

Therefore, we will mainly address the impact of the Omicron variant, which has been found to have a substantial impact on the IFR. The estimated hazard ratio of death

for Omicron variant compared with Delta variant in various studies ranges from 0.12 to 0.34. Here we set the relative risk of the Omicron variant compared to the pre-Omicron era as 0.25, which is roughly the average hazard ratio reported in Lewnard et al. (2022); Ulloa et al. (2022); Ward et al. (2022); Nyberg et al. (2022). Consequently, the relative risk $\mathrm{IFR}_O(t)$ is given by

$$\mathrm{IFR}_O(t) = (1 - O(t)) + 0.25O(t),$$

where $O(t)$ represents the proportion of the Omicron variant at time t. The time series $O(t)$ can be obtained through logistic regression analysis of sequencing data from Disease Control (2023c).

5.5 Calibration of State Baseline IFR

After acquiring information on how the IFR changes over time, age composition, vaccination rate, and variants, it is necessary to calibrate the baseline IFR using established results from modeling and serological surveys. The baseline IFR, denoted as IFR_b, represents the estimated IFR on July 1st, 2020. The time series of IFR is then expressed as

$$\mathrm{IFR}(t) = \mathrm{IFR}_b \times \mathrm{IFR}_r(t),$$

where $\mathrm{IFR}_r(t)$ has already been determined by combining all relevant factors.

We find that if the US average baseline IFR is applied to all states, the true case counts of a few states seems to be underestimated. To fix this issue, in this paper, we use two well-recognized studies published in Team (2022) and Irons and Raftery (2021) to calibrate our baseline IFR. The study published in Team (2022) utilizes serological surveys to estimate the IFR for each state on April 15, 2020, July 15, 2020, October 15, 2020, and January 1, 2021. On the other hand, the study described in Irons and Raftery (2021) employs a combination of modeling and serological surveys to estimate the IFR and undercounting factor (the ratio of true cases to confirmed cases) for each state on March 7, 2021. Both studies provide confidence intervals to account for the uncertainty in their estimates.

The method used to estimate IFR_b is as follows. We start by assuming $\mathrm{IFR}_b = 0.00754X$ to simplify the calculation, where 0.00754 represents the estimated IFR of the United States as of January 1st, 2021, as reported in Team (2022). The parameter X acts as a relative prefactor. Next, we use $\mathrm{IFR}_b = 0.00754$ to estimate the IFR on January 1st, 2021, March 7th, 2021, and the undercounting factor on March 7th, 2021. By comparing these estimated values with the data provided in Team (2022) and Irons and Raftery (2021), we can determine the likelihood of X for each state.

To estimate the probability density of X, we employ a Monte-Carlo-like approach. We assume that the two IFRs and the undercounting factor are normally distributed. The mean and variance of the normal distribution are derived from the estimated values and their respective confidence intervals. This approach yields an estimated probability density function for X. For example, if the IFR of a state on January 1st, 2021, using

the baseline IFR, is denoted as r_1, and the normal random variable representing the IFR of this state, based on Team (2022), has a mean of μ and a variance of σ^2, then this data suggests that X follows a normal probability density function $N(\mu/r_1, (\sigma/r_1)^2)$.

Let f_1, f_2, and f_3 denote the probability density functions obtained from the IFR in Team (2022), the IFR in Irons and Raftery (2021), and the undercounting factor in Irons and Raftery (2021), respectively. The likelihood of X is represented by the rescaled sum of these probability density functions, i.e., $f_1(x) + f_2(x) + f_3(x)$. The estimated baseline IFR, denoted as IFR_b, can then be calculated as

$$\text{IFR}_b = 0.00754 \times \frac{1}{3} \int_{-\infty}^{\infty} x(f_1(x) + f_2(x) + f_3(x))\mathrm{d}x,$$

which is equal to 0.00754 multiplied by the expectation of X. Additionally, the lower and upper bounds of IFR_b can be determined as the 0.05 and 0.95 percentiles of the rescaled probability density function $\frac{1}{3}(f_1(x) + f_2(x) + f_3(x))$, respectively.

In states with limited case counts or a significant number of COVID-19 cases in people of advanced age, the serological survey may tend to overestimate the IFR. This can result in the recovered true cases being smaller than the confirmed cases during certain time periods. To address this issue, we introduce an additional upper bound correction to ensure that the recovered true cases are not smaller than the confirmed cases. We use a criterion where, after a sufficient number of cases, the 50-day moving average of the recovered true cases should be greater than the confirmed cases. This provides an upper bound for the prefactor X.

If the calibrated X from the likelihood function exceeds the upper bound, we set X to be the upper bound value and adjust the lower bound of the confidence interval. The lower bound is reset to 0.6891 (which is the average ratio of the lower bound of the confidence interval to the estimated IFR in Team (2022)) multiplied by X. This additional correction is applied in a few states such as Virginia, Rhode Island, and Massachusetts.

Regarding the estimation of IFR in Vermont, there is a significant difference between the estimates in Team (2022) and Irons and Raftery (2021). We observe that the upper bound of X is close to the estimate in Irons and Raftery (2021). Taking into account the healthcare conditions, the estimated IFR for Vermont in Team (2022) (which is 2.498%) appears unreasonably high. Therefore, we only use the estimates from Irons and Raftery (2021) to calculate the likelihood of X for Vermont.

After calibration, we obtain the time series of IFR, denoted as Γ_{ifr}, for 10 selected states. These states are shown in Fig. 11. The time series of IFR for all states, including Washington DC, can be found in the Appendix.

After obtaining the time series of IFR, Γ_{ifr}, we proceed to solve the least square problem for each state, including Washington DC. The results for the 10 selected states are plotted in Fig. 12. The time series of all 51 states, spanning from February 29, 2020, to March 1, 2022, are provided in the Appendix. These 51 time series serve as the output data for the training set. It is evident that in most cases, the trend of the recovered true cases aligns with that of the confirmed cases.

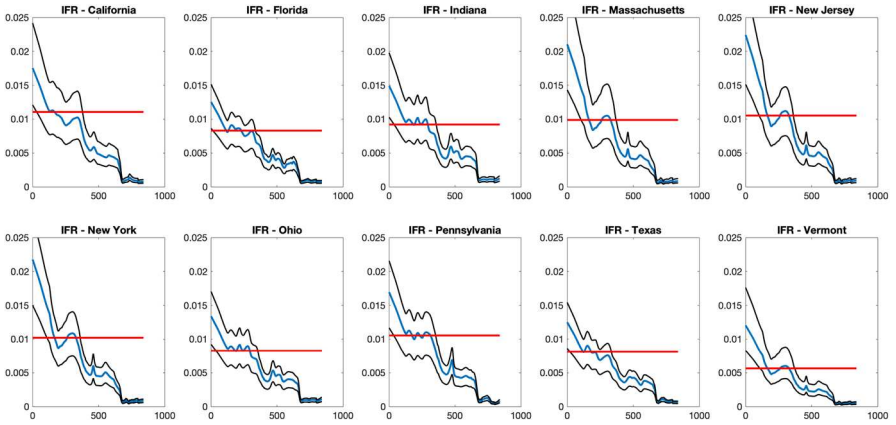

Fig. 11 Time series of each state after calibration (Color figure online)

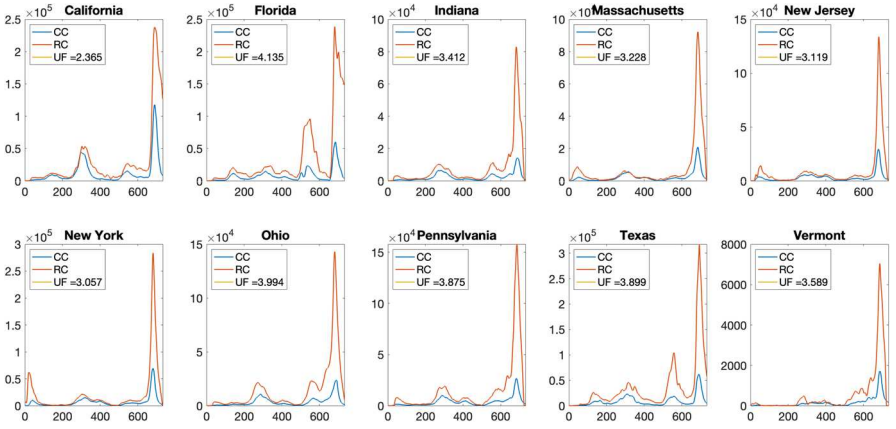

Fig. 12 Time series of recovered true case of 10 selected state and undercounting factor. CC: confirmed case. RC: recovered true case. UF: undercounting factor (Color figure online)

6 Data Source

In this section we summarize the data source used in our study to facilitate interested readers.

Daily confirmed cases count, daily death count, and state testing volume data are from the JHU COVID-19 database (Center 2023), available at https://github.com/CSSEGISandData/COVID-19

Infection case rate for each age group are from CDC (Disease Control 2023a), available at https://covid.cdc.gov/covid-data-tracker/#demographicsovertime

Vaccination rate are from CDC (Disease Control 2023b), available at https://data.cdc.gov/Vaccinations/COVID-19-Vaccination-Age-and-Sex-Trends-in-the-Uni/5i5k-6cmh

Vaccinated and unvaccinated incident rate ratio (IRR) of COVID-19 infection and death are from CDC (Disease Control 2023d), available at https://data.cdc.gov/Public-Health-Surveillance/Rates-of-COVID-19-Cases-or-Deaths-by-Age-Group-and/54ys-qyzm

Variant proportions data are from CDC (Disease Control 2023c), available at https://covid.cdc.gov/covid-data-tracker/#variant-proportions

7 Conclusion and Future Work

In this paper we use the backcasting method to estimate the true daily new case count of each state in the United States. The idea is that the true daily new case can be seen as a deconvolution problem because daily new death count is the convolution of a delay distribution and the product of the daily new case count and the infection fatality ratio (IFR). We first use time series of confirmed cases and deaths count from the whole United States to estimate the delay distribution from case to death. Then a significant portion of this paper is dedicated to estimating the IFR. The IFR can be inferred from serological survey and death count, both of which are considered as more trustworthy COVID-19 data. The time series of IFR depends one many additional factors, including age, vaccination, and viral variants. We use the case distribution of COVID-19 cases, weekly statistics of case and death reduction of vaccination, and daily composition of variants to estimate the time series of IFRs. The time series of IFR is then calibrated with some well-recognized studies that combines modeling and serological survey. While the effect of those factors on IFR are known to the community, as far as we aware, such a time series of IFR in the United States has not been estimated in other studies. The resulting estimated true case count is then used as the output training data for an artificial neural network to investigate the relation among testing volume, testing rate, confirmed case count, and true case count. Such relation is useful to public health agencies as it shows how many tests are necessary. It can also be use to provide a real-time true case count when the death count is not yet available or not reliable.

It is worth mentioning that the lack of reliable daily infection counts is a widespread problem in the area of COVID-19 modeling and prediction. Many modeling efforts resort to using death counts to infer model parameters. However, a fundamental assumption of a large class of compartment models is that the probability distribution for an individual transitioning from one compartment to another follows an exponential distribution. These compartment ODE models represent the infinite volume limit of continuous-time Markov chains (CTMC) that describe individual infections. The jump times of a CTMC must adhere to an exponential distribution because of the Markov property. However, as discussed in this paper and numerous other literature sources (Scheiner et al. 2020; Feng et al. 2007; Ghosh et al. 2022), the time from a confirmed case to a confirmed death significantly deviates from an exponential distribution. Therefore, assuming a linear transition rate from the infected population (I) to deaths (D) is problematic and leads to substantial deviations of the model from the real world.

Despite the progress made in this paper, we acknowledge that some additional work is necessary for our artificial neural network estimator to improve the accuracy

of estimate of the true daily new case count of COVID-19. One factor is that the home antigen test became widely available since 2022 late spring. As a result, a significant proportion of daily new COVID cases from low risk groups were not reported to the healthcare agency because many people just test themselves at home. This factor has not been addressed into our neural network estimator. In addition, today's IFR of COVID-19 is more difficult to estimate due to many factors. For example, the death count becomes less reliable due to higher immunity level and less virulent variants. Compared to the situation in 2020, it is less clear how many reported COVID-19 deaths are actually caused by COVID-19. One way of compensating this discrepancy is to consider the wastewater viral RNA data. Since late 2020, many cities and states test the COVID-19 RNA concentration in their wastewater regularly. It is known that wastewater surveillance is an important tool to infer the COVID-19 transmission dynamics (Shah et al. 2022; Daughton 2020). This serves as an invaluable bridge connecting the days when daily death count and IFR are more reliable (in 2020 and 2021) and the days when wastewater viral RNA data is available (after mid-2021). We expect the inclusion of wastewater data will provide a more accurate estimate of the COVID-19 daily case count in 2022 and afterward.

Acknowledgements We would like to thank REU students Ziyan Zhao for collecting vaccination data and Jessica Hu for collecting age group case data.

Author Contributions Yao Li and Ning Jiang are partially supported by NSF DMS-1813246 and DMS-2108628. Charles Kolozsvary is partially supported by the REU part of NSF DMS-1813246 and NSF DMS-2108628.

Data Availibility The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Appendix A: Additional Data About COVID-19

In this section we present many figures that demonstrate raw data, processed data, and intermediate results used to generate the training set. Some data for selected states have been already demonstrated in the main text. This includes

1 Time series of IFR for all 50 states plus Washington DC
2 Time series of recovered true cases and undercounting factor for all 50 states plus Washington DC
3 Raw and smoothed confirmed daily case count and daily death count for all 50 states plus Washington DC
4 Time series of case rate per age group at all regions of the United States
5 Time series of vaccination rate of all age group for all 50 states plus Washington DC
6 Incident rate ratio of COVID-19 case and death for vaccinated and unvaccinated groups.
7 Time series of testing volume for all 50 states plus Washington DC

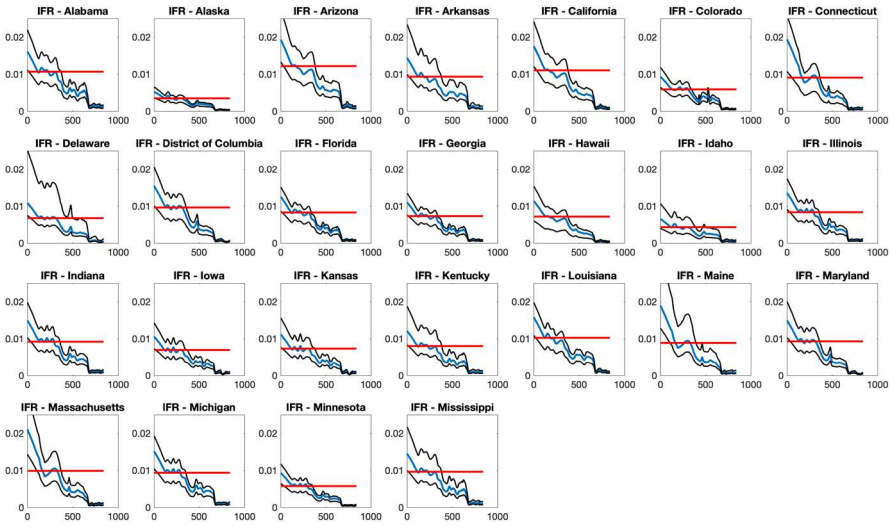

Fig. 13 Time series of state IFR for 24 states plus Washington DC (Color figure online)

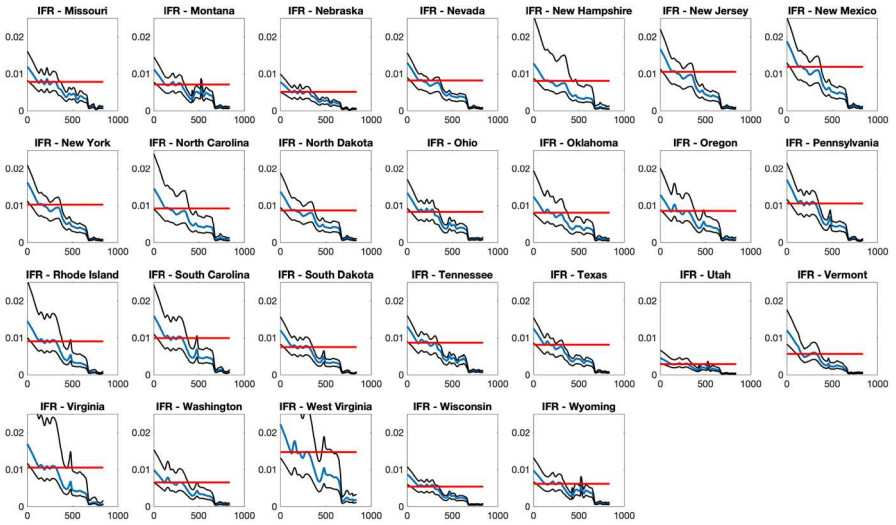

Fig. 14 Time series of state IFR for 26 states (Color figure online)

A.1 Time Series of State IFR

The time series of IFR for 10 selected states are presented in the main text. Below we demonstrate the time series of IFR for all 50 states plus Washington DC after considering age group case rate, vaccination, variant in Figs. 13 and 14.

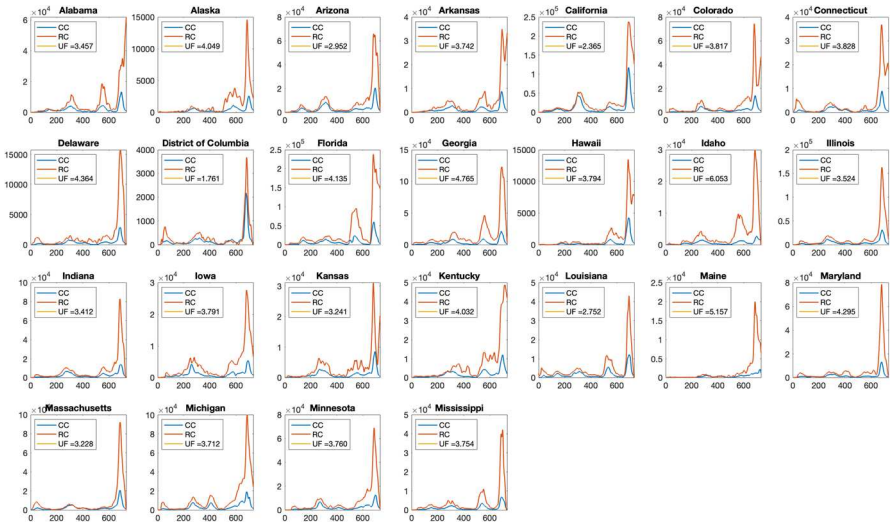

Fig. 15 Time series of recovered true case count for 24 states plus Washington DC (Color figure online)

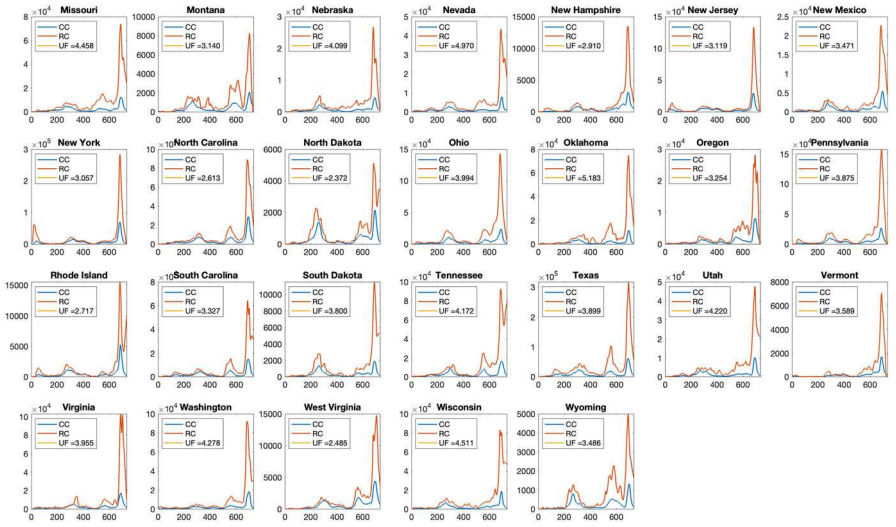

Fig. 16 Time series of recovered true case count for 26 states and Washington DC (Color figure online)

A.2 Time Series of State Recovered True Case

The time series of recovered true case and under counting factor for 10 selected states are demonstrated in the main text. Here we show these data for all 50 states plus Washington DC in Figs. 15 and 16.

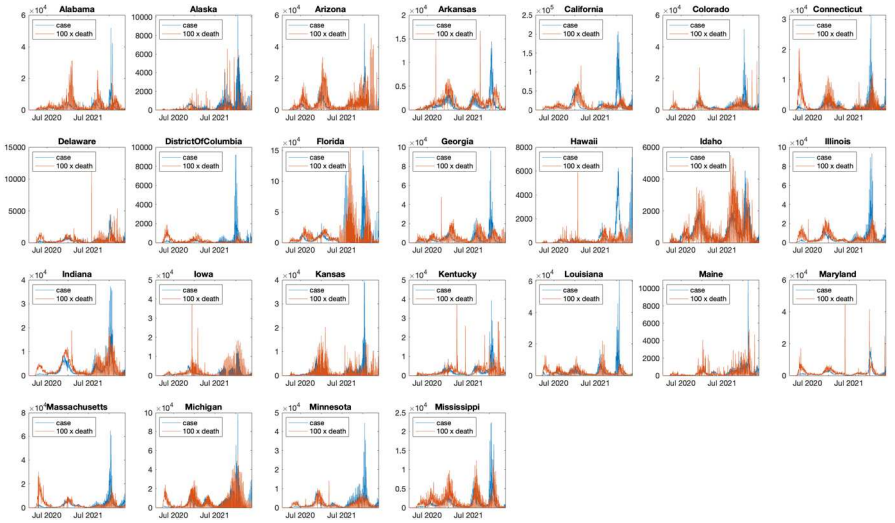

Fig. 17 Daily confirmed case count and 100× daily death count for 24 states plus Washington DC. Raw data before processing (Color figure online)

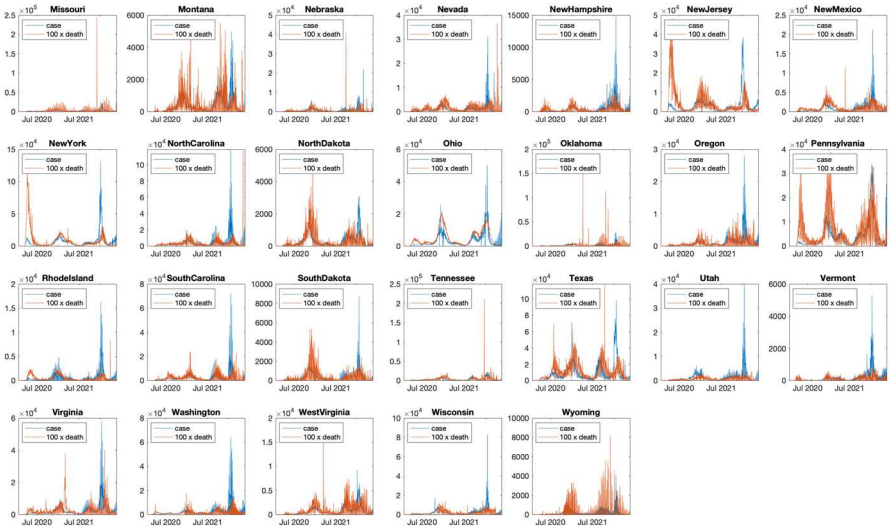

Fig. 18 Daily confirmed case count and 100× daily death count for 26 states. Raw data before processing (Color figure online)

A.3 State Confirmed Case and Death

Figures 17 and 18 show the daily case count and 100× daily death count of all 50 states plus Washington DC. The data comes from the JHU COVID-19 database (Center 2023). Figure 19 and 20 are the processed daily case count and daily death count after addressing data dump and holiday issues.

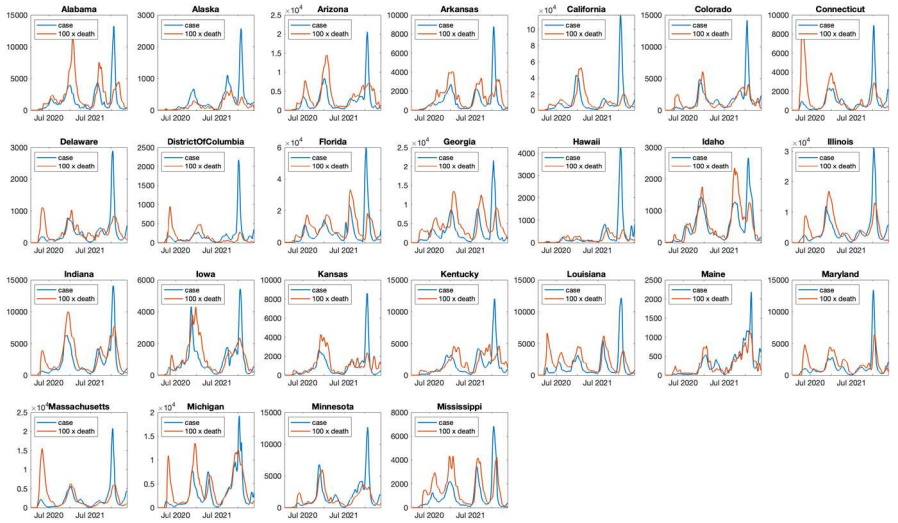

Fig. 19 Daily confirmed case count and 100× daily death count for 24 states plus Washington DC. Processed data after addressing weekday issue, holiday issue, and artificial data dump from backlogs (Color figure online)

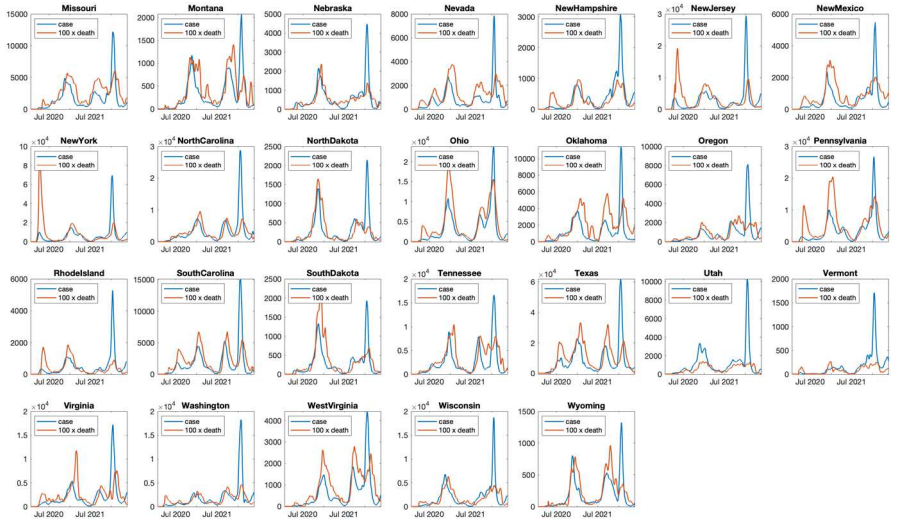

Fig. 20 Daily confirmed case count and 100× daily death count for 26 states. Processed data after addressing weekday issue, holiday issue, and artificial data dump from backlogs (Color figure online)

A.4 Case Rate Per Age Group

Figure 21 shows the time series of case rate of each age group from all 10 regions provided by CDC (Disease Control 2023a). The HHS regions used by CDC is described in the following Table 1.

Table 1 List of states and districts in each CDC region

Region	States
1	Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, Vermont
2	New Jersey, New York, Puerto Rico, Virgin Islands
3	Delaware, District Of Columbia, Maryland,Pennsylvania, Virginia, West Virginia
4	Alabama, Florida, Georgia, Kentucky, Mississippi, North Carolina, South Carolina, Tennessee
5	Illinois, Indiana, Michigan, Minnesota, Ohio, Wisconsin
6	Arkansas, Louisiana, New Mexico, Oklahoma, Texas
7	Iowa, Kansas, Missouri, Nebraska
8	Colorado, Montana, North Dakota, South Dakota, Utah, Wyoming
9	Arizona, California, Guam, Hawaii, Nevada
10	Alaska, Idaho, Oregon, Washington

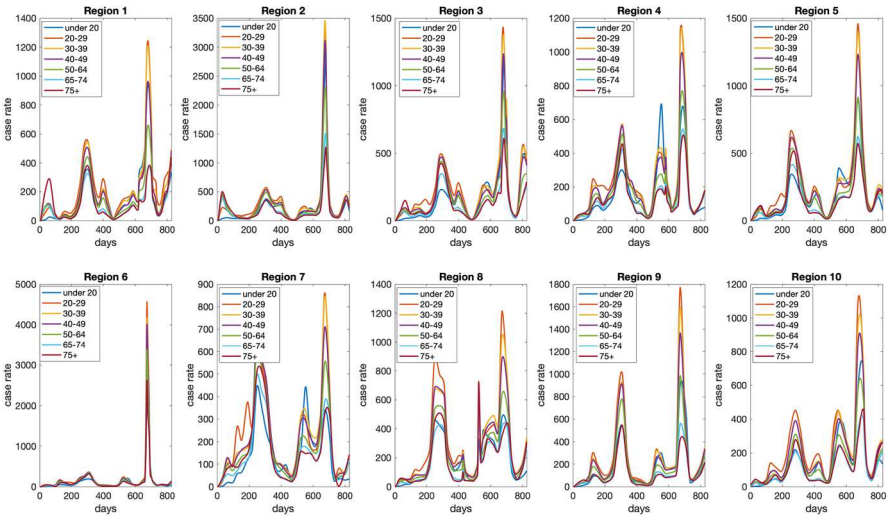

Fig. 21 Case rate of each age group in all 10 regions (Color figure online)

A.5 State Vaccination Rate

Figures 22 and 23 gives the time series of vaccinate rate for each age group older than 18 years old in all 50 states plus Washington DC. This data is obtained from CDC (Disease Control 2023b).

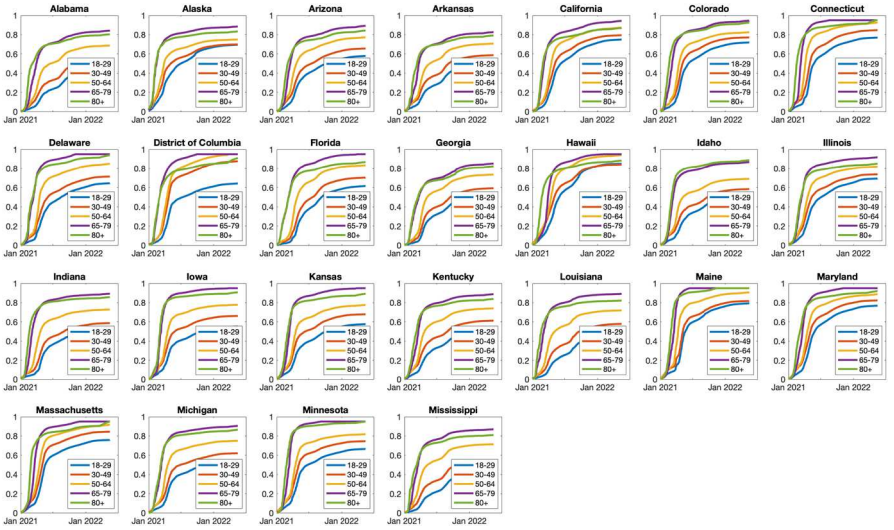

Fig. 22 Time series of vaccination rate of each age group for 24 states plus Washington DC (Color figure online)

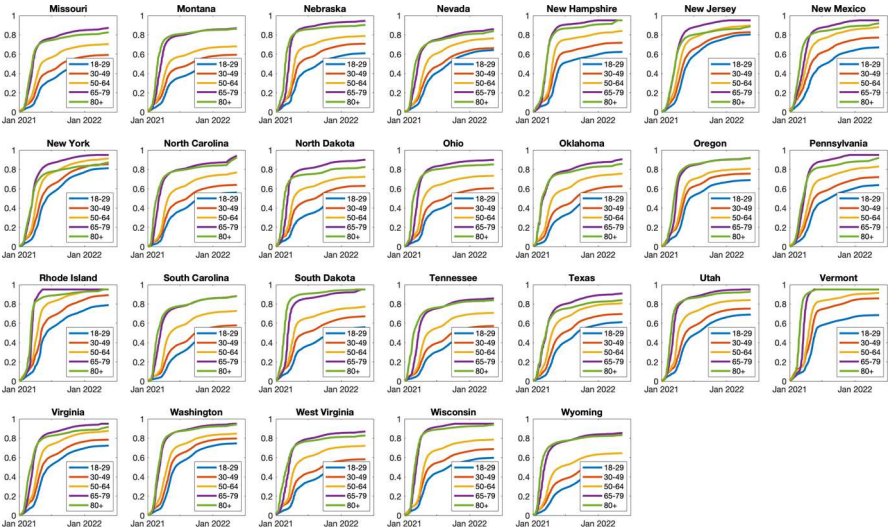

Fig. 23 Time series of vaccination rate of each age group for 26 states (Color figure online)

A.6 Incident Rate Ratio (IRR) of Vaccinated and Unvaccinated Groups

The incident ratio of COVID-19 infection and death for each group is given in Fig. 24. This data is obtained from CDC website (Disease Control 2023d). Note that death data of younger age group is not included because there are too few, sometimes zero, death count from vaccinated young group in many weeks. The ratio of IFR of unvaccinated group to vaccinated group of three older age groups are shown in Fig. 24 Right.

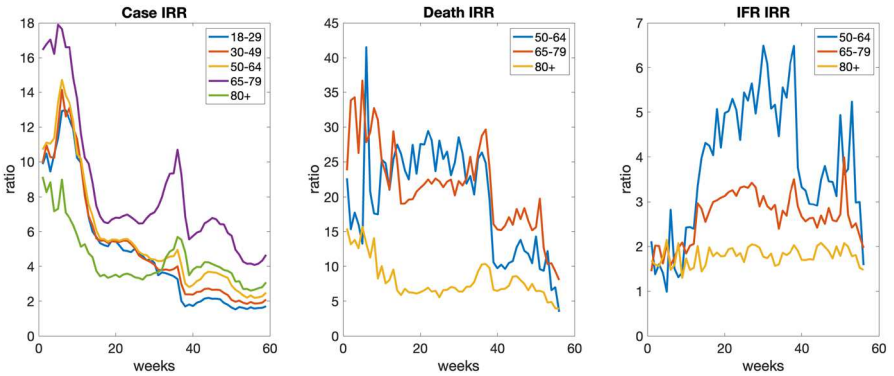

Fig. 24 Left and Middle: Incident rate ratio (IRR) of COVID-19 infection and death for each age group. Right: Ratio of IFR of unvaccinated group to vaccinated group (Color figure online)

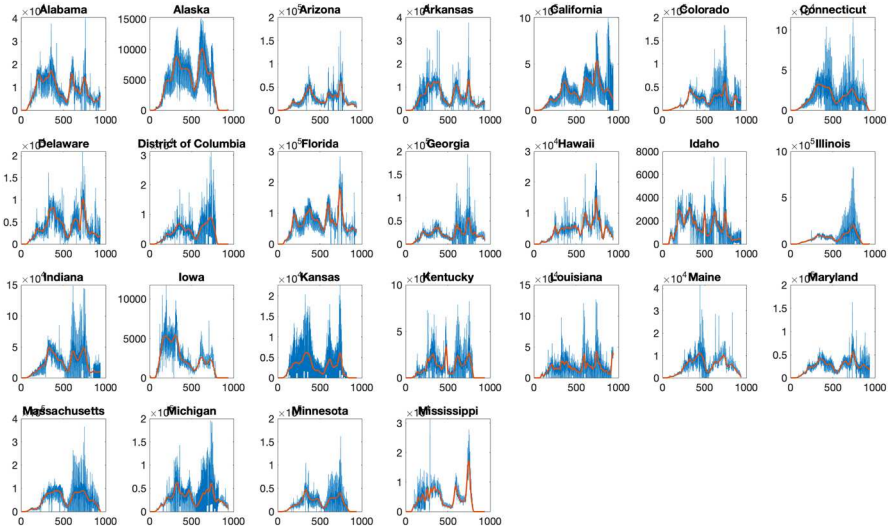

Fig. 25 Time series of COVID-19 testing volume for 24 states plus Washington DC. The blue and red lines are raw data and smoothed data respectively (Color figure online)

A.7 State Testing Volume

Figures 25 and 26 gives the time series of smoothed COVID-19 test volume in all 50 states plus Washington DC. This data comes from the Coronavirus Resource Center of Johns Hopkins University (Center 2023).

A.8 Training Set Data Distribution

Figure 27 displays the distribution of the data in the training data set.

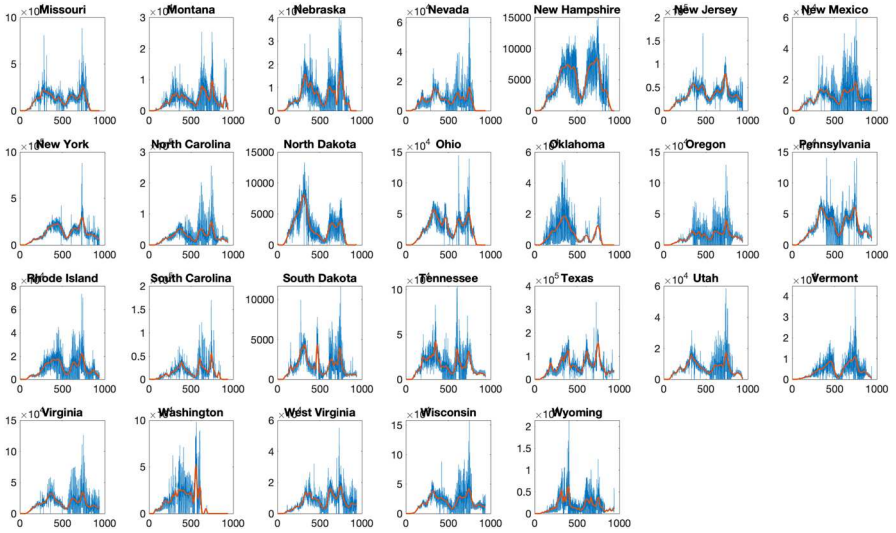

Fig. 26 Time series of COVID-19 testing volume for 26 states. The blue and red lines are raw data and smoothed data respectively (Color figure online)

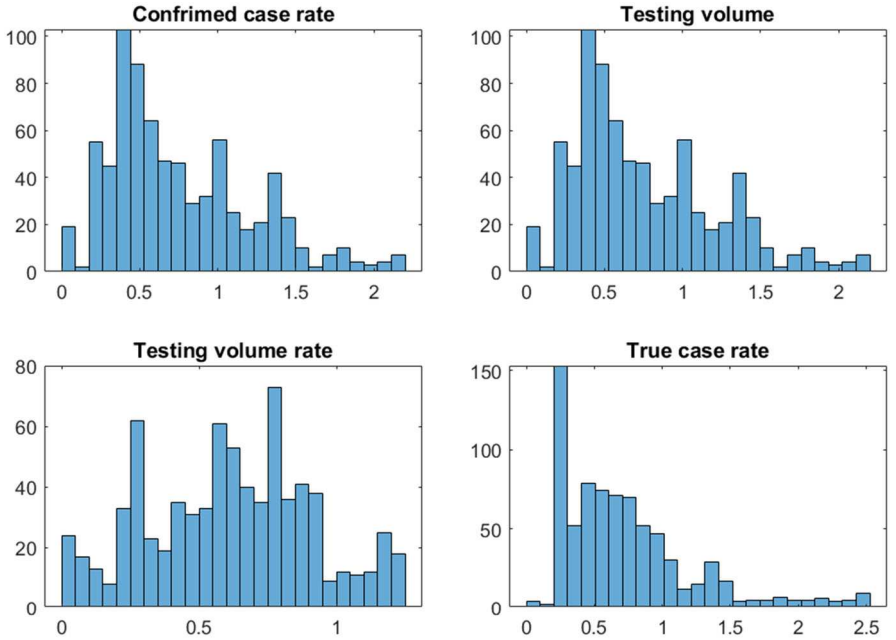

Fig. 27 Distribution of normalized data (Color figure online)

References

Akima H (1970) A new method of interpolation and smooth curve fitting based on local procedures. J ACM (JACM) 17(4):589–602

Akima H (1974) A method of bivariate interpolation and smooth surface fitting based on local procedures. Commun ACM 17(1):18–20

Albani V, Loria J, Massad E, Zubelli J (2021) Covid-19 underreporting and its impact on vaccination strategies. BMC Infect Dis 21:1–13

Barber RM, Sorensen RJ, Pigott DM, Bisignano C, Carter A, Amlag JO, Collins JK, Abbafati C, Adolph C, Allorant A (2022) Estimating global, regional, and national daily and cumulative infections with sars-cov-2 through Nov 14, 2021: a statistical analysis. Lancet 399(10344):2351–2380

Brazeau NF, Verity R, Jenks S, Fu H, Whittaker C, Winskill P, Dorigatti I, Walker PG, Riley S, Schnekenberg RP (2022) Estimating the covid-19 infection fatality ratio accounting for seroreversion using statistical modelling. Commun Med 2(1):54

Center JHCR (2023) COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. https://github.com/CSSEGISandData/COVID-19. Accessed 10 June 2023

Chen Z, Feng L, Lay HA Jr, Furati K, Khaliq A (2022) Seir model with unreported infected population and dynamic parameters for the spread of covid-19. Math Comput Simul 198:31–46

Chimmula VKR, Zhang L (2020) Time series forecasting of covid-19 transmission in Canada using lstm networks. Chaos Solitons Fractals 135:109864

Cleveland WS, Devlin SJ (1988) Locally weighted regression: an approach to regression analysis by local fitting. J Am Stat Assoc 83(403):596–610

Dairi A, Harrou F, Zeroual A, Hittawe MM, Sun Y (2021) Comparative study of machine learning methods for covid-19 transmission forecasting. J Biomed Inform 118:103791

Daughton CG (2020) Wastewater surveillance for population-wide covid-19: the present and future. Sci Total Environ 736:139631

Disease Control C (2023a) Prevention: COVID-19 weekly cases and deaths per 100,000 population by age, race/ethnicity, and sex. https://covid.cdc.gov/covid-data-tracker/#demographicsovertime. Accessed 10 June 2023

Disease Control C (2023b) Prevention: COVID-19 Vaccination Age and Sex Trends in the United States, National and Jurisdictional. https://data.cdc.gov/Vaccinations/COVID-19-Vaccination-Age-and-Sex-Trends-in-the-Uni/5i5k-6cmh. Accessed 10 June 2023

Disease Control C (2023c) Prevention: COVID data tracker: Variant Proportion. https://covid.cdc.gov/covid-data-tracker/#variant-proportions. Accessed 10 June 2023

Disease Control C (2023d) Prevention: Rates of COVID-19 Cases and Deaths by Vaccination Status. https://data.cdc.gov/Public-Health-Surveillance/Rates-of-COVID-19-Cases-or-Deaths-by-Age-Group-and/54ys-qyzm. Accessed 10 June 2023

Dutta R, Das N, Majumder M, Jana B (2023) Aspect based sentiment analysis using multi-criteria decision-making and deep learning under covid-19 pandemic in India. CAAI Trans Intell Technol 8(1):219–234

Easton DM, Hirsch HR (2008) For prediction of elder survival by a Gompertz model, number dead is preferable to number alive. Age 30:311–317

Feng Z, Xu D, Zhao H (2007) Epidemiological models with non-exponentially distributed disease stages and applications to disease control. Bull Math Biol 69(5):1511–1536

Flaxman S, Mishra S, Gandy A, Unwin H, Coupland H, Mellan T, Zhu H, Berah T, Eaton J, Perez Guzman P, et al (2020) Report 13: Estimating the number of infections and the impact of non-pharmaceutical interventions on covid-19 in 11 european countries

Ghosh S, Volpert V, Banerjee M (2022) An epidemic model with time-distributed recovery and death rates. Bull Math Biol 84(8):78

Guo Q, He Z (2021) Prediction of the confirmed cases and deaths of global covid-19 using artificial intelligence. Environ Sci Pollut Res 28:11672–11682

Harvey WT, Carabelli AM, Jackson B, Gupta RK, Thomson EC, Harrison EM, Ludden C, Reeve R, Rambaut A, Consortium C-GUC-U (2021) Sars-cov-2 variants, spike mutations and immune escape. Nat Rev Microbiol 19(7):409–424

He S, Peng Y, Sun K (2020) Seir modeling of the covid-19 and its dynamics. Nonlinear Dyn 101:1667–1680

Hortaçsu A, Liu J, Schwieg T (2021) Estimating the fraction of unreported infections in epidemics with a known epicenter: An application to covid-19. J Econom 220(1):106–129

Irons NJ, Raftery AE (2021) Estimating sars-cov-2 infections from deaths, confirmed cases, tests, and random surveys. Proc Natl Acad Sci 118(31):2103272118. https://doi.org/10.1073/pnas.2103272118
Jahja M, Chin A, Tibshirani RJ (2022) Real-time estimation of covid-19 infections: deconvolution and sensor fusion. Stat Sci 37(2):207–228
Jahja M, Chin A, Tibshirani RJ (2022) Real-time estimation of COVID-19 infections: deconvolution and sensor fusion. Stat Sci 37(2):207–228. https://doi.org/10.1214/22-STS856
Kamalov F, Rajab K, Cherukuri AK, Elnagar A, Safaraliev M (2022) Deep learning for covid-19 forecasting: State-of-the-art review. Neurocomputing 511:142–154
Kevrekidis GA, Rapti Z, Drossinos Y, Kevrekidis PG, Barmann MA, Chen Q-Y, Cuevas-Maraver J (2022) Backcasting covid-19: a physics-informed estimate for early case incidence. R Soc Open Sci 9(12):220329
Kidger P, Lyons T (2020) Universal approximation with deep narrow networks. In: Conference on learning theory. PMLR, pp 2306–2327
Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. arXiv:1412.6980
Lewnard JA, Hong VX, Patel MM, Kahn R, Lipsitch M, Tartof SY (2022) Clinical outcomes associated with sars-cov-2 omicron (b. 1.1. 529) variant and ba. 1/ba. 1.1 or ba. 2 subvariant infection in Southern California. Nat Med 28(9):1933–1943
Maiorov V, Pinkus A (1999) Lower bounds for approximation by mlp neural networks. Neurocomputing 25(1–3):81–91
Meyerowitz-Katz G, Merone L (2020) A systematic review and meta-analysis of published research data on covid-19 infection fatality rates. Int J Infect Dis 101:138–148
Miller AC, Hannah LA, Futoma J, Foti NJ, Fox EB, D'Amour A, Sandler M, Saurous RA, Lewnard JA (2022) Statistical deconvolution for inference of infection time series. Epidemiology (Cambridge, Mass.) 33(4):470
Miller AC, Hannah L, Futoma J, Foti NJ, Fox EB, D'Amour A, Sandler M, Saurous RA, Lewnard JA (2022) Statistical deconvolution for inference of infection time series. Epidemiology 33(4):470–479. https://doi.org/10.1097/EDE.0000000000001495
Namasudra S Dhamodharavadhani S, Rathipriya R (2021) Nonlinear neural network based forecasting model for predicting covid-19 cases. Neural Process Lett 1–21
Nyberg T, Ferguson NM, Nash SG, Webster HH, Flaxman S, Andrews N, Hinsley W, Bernal JL, Kall M, Bhatt S (2022) Comparative analysis of the risks of hospitalisation and death associated with sars-cov-2 omicron (b. 1.1. 529) and delta (b. 1.617. 2) variants in England: a cohort study. Lancet 399(10332):1303–1312
Organization WH (2023) WHO Coronavirus (COVID-19) Dashboard. https://covid19.who.int/. Accessed 10 June 2023
Perc M, Gorišek Miksić N, Slavinec M, Stožer A (2020) Forecasting covid-19. Front Phys 8:127
Phipps SJ, Grafton RQ, Kompas T (2020) Robust estimates of the true (population) infection rate for covid-19: a backcasting approach. R Socy Open Sci 7(11):200909. https://doi.org/10.1098/rsos.200909
Rahimi I, Chen F, Gandomi AH (2023) A review on covid-19 forecasting models. Neural Comput Appl 35(33):23671–23681
Raissi M, Perdikaris P, Karniadakis GE (2019) Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. J Comput Phys 378:686–707
Sarría-Santamera A, Abdukadyrov N, Glushkova N, Russell Peck D, Colet P, Yeskendir A, Asúnsolo A, Ortega MA (2022) Towards an accurate estimation of covid-19 cases in Kazakhstan: back-casting and capture-recapture approaches. Medicina 58(2):253
Scheiner S, Ukaj N, Hellmich C (2020) Mathematical modeling of covid-19 fatality trends: death kinetics law versus infection-to-death delay rule. Chaos Solitons Fractals 136:109891
Shah S, Gwee SXW, Ng JQX, Lau N, Koh J, Pang J (2022) Wastewater surveillance to infer covid-19 transmission: a systematic review. Sci Total Environ 804:150060
Tang S, Cao Y (2023) A phenomenological neural network powered by the national wastewater surveillance system for estimation of silent covid-19 infections. Sci Total Environ 902:166024
Team C-F (2022) Variation in the covid-19 infection-fatality ratio by age, time, and geography during the pre-vaccine era: a systematic analysis. Lancet 399(10334):1469–1488 https://doi.org/10.1016/S0140-6736(21)02867-1
Ulloa AC, Buchan SA, Daneman N, Brown KA (2022) Estimates of sars-cov-2 omicron variant severity in Ontario, Canada. JAMA 327(13):1286–1288

Vaid S, Cakan C, Bhandari M (2020) Using machine learning to estimate unobserved covid-19 infections in North America. J Bone Joint Surg. American volume

Ward IL, Bermingham C, Ayoubkhani D, Gethings OJ, Pouwels KB, Yates T, Khunti K, Hippisley-Cox J, Banerjee A, Walker AS, et al (2022) Risk of covid-19 related deaths for sars-cov-2 omicron (b. 1.1. 529) compared with delta (b. 1.617. 2): retrospective cohort study. bmj 378

Watson GL, Xiong D, Zhang L, Zoller JA, Shamshoian J, Sundin P, Bufford T, Rimoin AW, Suchard MA, Ramirez CM (2021) Pandemic velocity: Forecasting covid-19 in the us with a machine learning & Bayesian time series compartmental model. PLoS Comput Biol 17(3):1008837

Wu SL, Mertens AN, Crider YS, Nguyen A, Pokpongkiat NN, Djajadi S, Seth A, Hsiang MS, Colford JM Jr, Reingold A (2020) Substantial underestimation of sars-cov-2 infection in the united states. Nat Commun 11(1):4507

Zhai J, Dobson M, Li Y (2022) A deep learning method for solving Fokker–Planck equations. In: Mathematical and scientific machine learning. PMLR, pp 568–597