frontiers | Frontiers in Public Health

TYPE Original Research
PUBLISHED 26 June 2024
DOI 10.3389/fpubh.2024.1359368

Accuracy of US CDC COVID-19 forecasting models

Aviral Chharia[1,2,3†‡], Govind Jeevan[1,2‡], Rajat Aayush Jha[1,2†‡], Meng Liu[1,4‡], Jonathan M. Berman[1,5] and Christin Glorioso[1,2,6*]

[1]Global Health Research Collective, Academics for the Future of Science, Cambridge, MA, United States, [2]Data Informatics Center for Epidemiology, PathCheck Foundation, Cambridge, MA, United States, [3]Department of Mechanical Engineering, Thapar Institute of Engineering and Technology, Patiala, PB, India, [4]Department of Industrial and Manufacturing Engineering, Penn State University, University Park, PA, United States, [5]Department of Basic Science, New York Institute of Technology, College of Osteopathic Medicine at Arkansas State University, Jonesboro, AR, United States, [6]Department of Anatomy, University of California, San Francisco, San Francisco, CA, United States

OPEN ACCESS

EDITED BY
Tamer Oraby,
The University of Texas Rio Grande Valley,
United States

REVIEWED BY
Sofia Seabra,
New University of Lisbon, Portugal
Junxiang Chen,
Indiana University, United States

*CORRESPONDENCE
Christin Glorioso
✉ gloriosoca@gmail.com

†PRESENT ADDRESSES
Aviral Chharia,
Department of Mechanical Engineering,
Carnegie Mellon University, Pittsburgh, PA,
United States

Rajat Aayush Jha,
School of Computing and Augmented
Intelligence, Ira A. Fulton Schools of
Engineering, Arizona State University, Tempe,
AZ, United States

‡These authors have contributed equally to this work and share first authorship

RECEIVED 16 January 2024
ACCEPTED 07 June 2024
PUBLISHED 26 June 2024

CITATION
Chharia A, Jeevan G, Jha RA, Liu M, Berman JM and Glorioso C (2024) Accuracy of US CDC COVID-19 forecasting models. *Front. Public Health* 12:1359368. doi: 10.3389/fpubh.2024.1359368

Accurate predictive modeling of pandemics is essential for optimally distributing biomedical resources and setting policy. Dozens of case prediction models have been proposed but their accuracy over time and by model type remains unclear. In this study, we systematically analyze all US CDC COVID-19 forecasting models, by first categorizing them and then calculating their mean absolute percent error, both wave-wise and on the complete timeline. We compare their estimates to government-reported case numbers, one another, as well as two baseline models wherein case counts remain static or follow a simple linear trend. The comparison reveals that around two-thirds of models fail to outperform a simple static case baseline and one-third fail to outperform a simple linear trend forecast. A wave-by-wave comparison of models revealed that no overall modeling approach was superior to others, including ensemble models and errors in modeling have increased over time during the pandemic. This study raises concerns about hosting these models on official public platforms of health organizations including the US CDC which risks giving them an official imprimatur and when utilized to formulate policy. By offering a universal evaluation method for pandemic forecasting models, we expect this study to serve as the starting point for the development of more accurate models.

KEYWORDS

public health interventions, biomedical engineering, machine learning, COVID-19, time series forecasting, pandemics

1 Introduction

The COVID-19 pandemic (1) resulted in at least 100 million confirmed cases and more than 1 million deaths in the United States alone. Worldwide, cases exceed 650 million, with at least 6.5 million deaths (2). The pandemic has affected every country and presented a major threat to global health. This has caused a critical need to study the transmission of emerging infectious diseases in order to make accurate case forecasts, especially during disease outbreaks.

1.1 Background

Case prediction models are useful for developing pandemic preventive and control methods, such as suggestions for healthcare infrastructure needs, isolation of infected persons, and contact activity tracking. Accurate models can allow better decision-making about the degree of precautions necessary for a given region at a particular time, which regions to avoid travel to, and the degree of risk in various activities like public gatherings. Likewise, models can be used to proactively prepare for severe surges in cases by allocating biomedical resources such as oxygen or personnel. Collecting and presenting these models gives public health officials, and organizations such as the UNITED STATES CENTERS FOR DISEASE CONTROL AND PREVENTION (US CDC) (3), a mechanism to disseminate these predictions to the public, but risks giving them an official imprimatur, suggesting that these models were either developed or endorsed by a government agency.

1.2 Motivation

Since the start of the pandemic, dozens of case prediction models for the US have been designed using a variety of methods. Each of these models depends on data available about cases, derived from a heterogeneous system of reporting, which can vary by county and suffer from regional and temporal delays. For example, some counties may collect data over several days and make it public at once, which creates an illusion of a sudden burst of cases. In counties with less robust testing programs, the lack of data can limit modeling accuracy. These methods are not uniform or standardized between groups that perform data collection, resulting in unpredictable errors. Underlying biases in the data, such as under-reporting, can produce predictable errors in model quality, requiring models to be adjusted to predict future erroneous reporting rather than actual case numbers. Such under-reporting has been identified by serology data (4, 5). Moreover, there is no universally agreed upon system for assessing and comparing the accuracy of case prediction models. Often published models use different methods, which makes direct comparisons difficult. The CDC has taken in data of case prediction models in a standardized way which makes direct comparisons possible (3). In this study, we use Mean Absolute Percent Error (MAPE) compared to the true case numbers to compare models for normalization purposes. First, we consider which models have the most accurately predicted case counts with the least MAPE. Next, we divide these models into five broad sub-types based on approach, i.e., epidemiological (or compartment) models, machine learning approaches, ensemble approaches, hybrid approaches, and other approaches, and compare the overall error of models using these approaches. We also consider which exclusion criteria might produce ensemble models with the greatest accuracy and predictive power.

1.3 Contributions

A few studies (6, 7) have compared COVID-19 case forecasting models. However, the present study is unique in several aspects. First, it is focused on prediction models of US cases and takes into consideration all CDC models that pass the set inclusion criterion. Second, since these models were uploaded in a standardized format they can be compared across several dimensions such as R_0, peak timing error, percent error, and model architecture. Third, we seek to answer several unaddressed questions relevant to pandemic case modeling:

- First, can we establish a metric to uniformly evaluate pandemic forecasting models?
- Second, What are the top-performing models during the four COVID-19 waves in the US and how do these fare on the complete timeline?
- Third, are there categories or classes of case prediction models that perform significantly better than others?
- Fourth, how do model predictions fare with increased forecast horizons?
- Lastly, how do models compare to two simple baselines?

2 Methodology

This work compares various US-CDC COVID-19 forecasting models by their quantitative aspects evaluating their performance in strictly numerical terms over various time segments. The US-CDC collects weekly forecasts for COVID cases in four different horizons: 1-, 2-, 3-, and 4-weeks, i.e., each week, the models make a forecast for new COVID cases in each of the four consecutive weeks from the date of the forecast. The forecast horizon is defined as the length of time into the future for which forecasts are to be prepared. In the present study, we focus on evaluating the performance of models for their 4-week ahead forecasts.

2.1 Datasets

The data for the confirmed case counts are taken from the COVID-19 Data Repository (https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_daily_reports_us), maintained by the Center for Systems Science and Engineering at the Johns Hopkins University. The data for the predicted case counts, of all the models, is obtained from the data repository for the COVID-19 Forecast Hub (https://github.com/reichlab/covid19-forecast-hub), which is also the data source for the official US-CDC COVID-19 forecasting page. Both these datasets were pre-processed to remove models that have made predictions for <25% of the target dates covered by the respective time segment. For plotting Figure 1A, the PYPLOT module from the `Matplotlib` library in Python was used.

2.2 Categorizing models

The models were categorized into five different categories- Epidemiological/ Compartmental (see Table 1), Machine Learning (see Table 2), Ensemble (see Table 3), Hybrid (see Table 4), and Others (see Table 5). This was based on keywords found in model names and going to each model description on their respective web pages and articles. The models which did not broadly fall into these categories were kept in "others". These models use very different methods to arrive at predictions. We comprehensively analyzed 51 models. The CDC also uses an

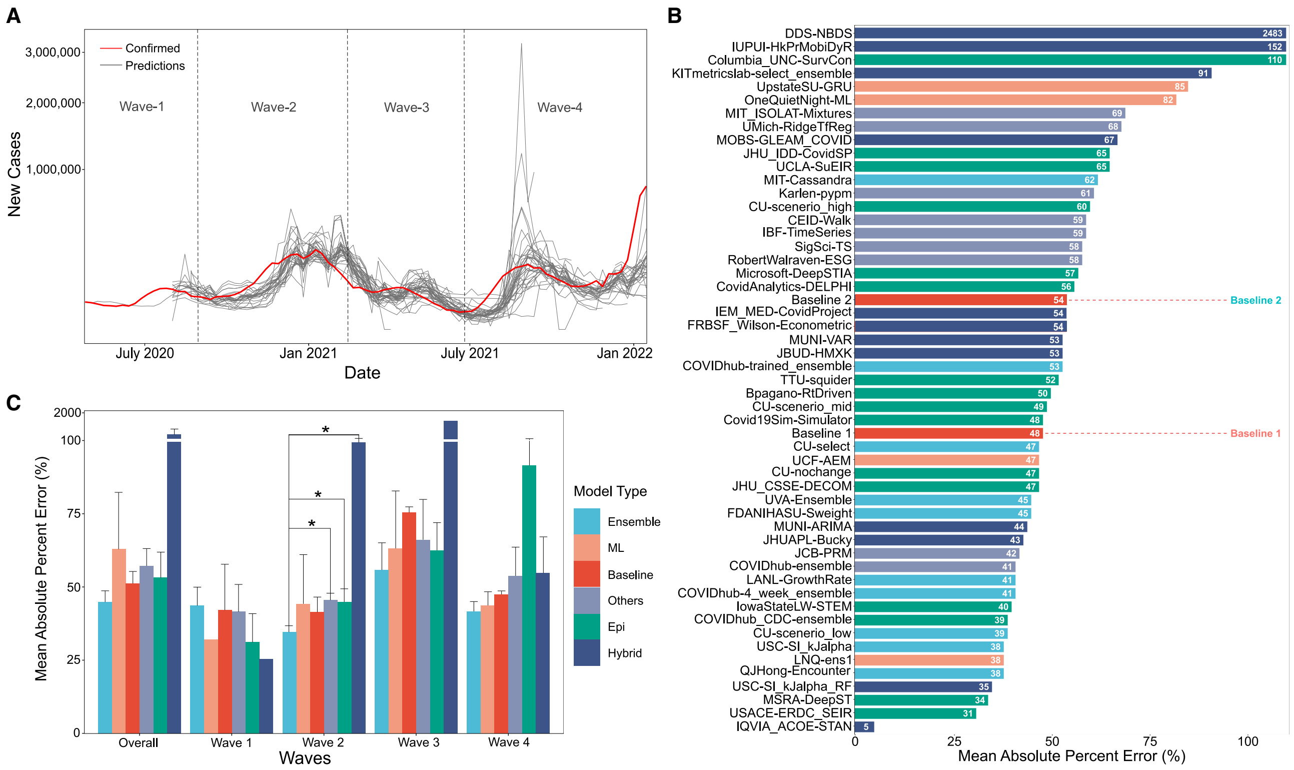

FIGURE 1
(A) Visual overlay of real case counts and predicted case counts across all waves examined. Actual case counts are shown in red, predicted counts are shown in gray with each trace representing a different CDC COVID-19 forecasting model on the sqrt plot. **(B)** MAPE values of US-CDC case prediction models on the complete timeline, i.e., Wave-I to IV. The *y*-axis is sorted descending from lowest error to highest. The color scheme represents the model category. **(C)** Bar graph showing the non-parametric Kruskall-Wallis test results. A Mann-Whitney test was further performed for groups with significant differences. Category-wise error was achieved by the models both overall and wave-wise from wave-1 to wave-4. Note that Hybrid models have a high MAPE, i.e., overall: 261.16%, Wave 1: 25.36%, Wave 4: 99.3%, Wave 3: 1421.325%, Wave 4: 54.74%).

ensemble model, and we looked at whether this was better than any individual model. For each model uploaded to the CDC website, MAPE was calculated and reported in this study, and the models were compared wave-wise as shown in various figures. For each model, the model type was noted, as well as the month proposed.

2.3 Wave definition

A thorough search reveals M.D., epidemiologists, and policymakers do share the same underlying principles of the term “wave". Popular media explain that “the word ‘wave’ implies a natural pattern of peaks and valleys”. WHO stated in order to say one wave is ended, the virus has to be brought under control and cases have to fall substantially, then for a second wave to start, you need a sustained rise in infections. As part of their National Forecasts for COVID cases, the CDC has reported the results from a total of 54 different models at various instances of time during the pandemic. We define the waves, i.e., Wave-I: July, 6th 2020 to August 31st, 2020, Wave-II: September, 1st 2020 to February, 14th 2021, Wave-III: February, 15th 2021 to July, 26th 2021 and Wave-IV: July, 27th 2021 to January, 17th 2022, corresponding to each of the major waves in the US.

2.4 Baselines

The performance of the models was evaluated against two simple baselines. Baseline-I is the “CovidHub-Baseline" (or CDC’s baseline), i.e., the median prediction at all future horizons is the most recent observed incidence (i.e., the most recent day). Baseline-II is the linear predictor extrapolation using the slope of change in reported active cases between the 2 weeks preceding the date of the forecast. These baselines are included in the bar charts (shown in Figures 1B, 2). Within each of the waves of interest, only models that made a significant number of predictions are considered for comparison. We only consider models that have made predictions for at least 25% of the target dates covered by the respective time segment for all comparisons in this section. The MAPE was calculated on the 4-week forecast horizon. Figure 1B illustrates the performance of models across all the waves. Same procedure, as in Figure 2, was followed for plotting Figure 1B.

2.5 Model comparison

The performance of all the models is compared, wave-wise and on the complete timeline based on the MAPE (or mean absolute percent error). MAPE is defined as the ratio of absolute percentage

TABLE 1 Epidemiological/compartmental US-CDC COVID-19 forecasting models—their description, features employed, methodology and assumptions they make regarding public health interventions.

Model	Features used for forecasting	Author	Method	Assumptions
TTU-Squider (8)	Takes into account power-law incident rate, separate compartments for silent spreaders, quarantine/hospital isolation of infected individuals, social contact restrictions, possible loss of immunity for recovered individuals.	Hussain Lab, Texas Tech University	SIR	Effects of interventions are reflected in observed data and will continue going forward.
JHU-IDD (9)	Accounts for uncertainty in epidemiological parameters including R0, spread of more transmissible variants, infectious period, time delays to health outcomes and effectiveness of state-wide intervention policies.	JHU IDD Working Group	Meta population SEIR	Current interventions will not change during the period forecasted.
IowaStateLW-STEM (10)	A non-parametric space-time disease transmission model for epidemic data to study the spatial-temporal pattern of COVID-19.	Iowa State- Lily Wang's Group	Non-parametric spatiotemporal model	—
BPagano-RtDriven (11)	The effective transmission ratio, Rt, drives the model's projections. To forecast how Rt will change with time, the model analyzes Rt change data through the pandemic and applies a model of that characteristic behavior to forecast infections.	BPagano	SIR	Effects of interventions are reflected in observed data and will continue going forward.
UCLA-SuEIR (12)	An SEIR Model variant that takes into consideration the effects of re-openings. Assumes a transition from a virtual "Quarantined" group to the "Susceptible" group at a specific rate for the states that have reopened/ partially reopened. Most notable feature is that it can infer untested cases as well as unreported cases.	UCLA Statistical Machine Learning Lab	Modified SEIR	Contact rates will increase as states reopen and calculate the increase in contact rates for each state.
COVID19Sim-Simulator (13)	Uses a validated compartment model defined using SEIR with continuous-time progression to simulate the trajectory of COVID-19 at the state level.	COVID-19 Simulator	SEIR	Based on assumptions about how in the future, the levels of social distancing may evolve.
USACE-ERDC_SEIR (14)	Bayesian Inference calculates model parameters from observations of total number of cases. A prior probability distribution over the model parameters. The accumulated observations & subject matter knowledge are then coupled with a statistical model of model-data mismatch to generate a posterior probability distribution across model parameters. To make forecasts, parameters maximizing posterior probability density are used.	US Army Engineer Research & Development Center	Process-based classic SEIR model with compartments for unreported infections/ isolated individuals.	(i) Current interventions don't change during forecast period. (ii) Modeled populations are large enough that disease states fluctuations grow slower than average. (iii) Recovered individuals are not infectious/ susceptible to infections.
Microsoft-DeepSTIA (15)	Deep Spatio-temporal network with intervention under the assumption of Spatio-temporal process in the pandemic of different regions.	Microsoft	SEIR model on spatiotemporal network	Current interventions will not change during the period forecasted.
CovidAnalytics-DELPHI (16)	Introduces new states to accommodate for unnoticed cases, as well as an explicit death state. A non-linear curve reflecting government reaction is used to adjust the infection rate. Also, a meta-analysis of 150 factors is used to determine key illness parameters, while epidemiological parameters are fitted to historical death counts & identified cases.	MIT Covid-Analytics	Augmentation of SEIR model	
Columbia_UNC-SurvCon (17)	Considers transmission throughout pre-symptomatic incubation phase, employing a time-varying effective R0 to capture the temporal trend of transmission & change in response to a public health intervention. Uses permutation to quantify uncertainty.	Columbia_UNC	—	—
CU-select, CU-nochange, CU-scenario_low, CU-scenario_mid, CU-scenario_high (18)	Produces different intervention scenarios, each assuming various interventions & rates of compliance are implemented in the future. (i) Presents the weekly scenario believed to be most plausible given current observations & planned intervention policies. (ii) Current contact rates will remain unchanged in the future. Assumes relatively (iii) low transmission, (iv) moderate transmission, & (v) high transmission	Columbia University	Metapopulation county-level SEIR	

errors of the predictions normalized by the number of data points. The error refers to the difference between the confirmed case counts and the predicted case counts and is calculated as shown in Equation (1) below:

$$MAPE = (1/n) \sum_{t=1}^{n} |(A_t - P_t)/A_t| \tag{1}$$

Here, n = number of data points, A_t is the actual value and P_t denoted the predicted value.

Figure 1C shows the on-parametric Kruskall-Wallis test results on the category-wise errors, achieved by the models overall as well as wave-wise. The mean of the MAPE values is calculated for each category in model type, for overall and each wave separately. The mean is calculated by adding the MAPE values of all the models in a category and dividing it by the number of models in that category for the corresponding wave. hen, a non-parametric Kruskall-Wallis test is performed to determine if there is a significant difference between the two groups. We used `scipy` in Python to perform the test. For cases where significant differences were found (for Wave-II), a Mann-Whitney test was further performed. We did not perform the popular ANOVA test since ANOVA conditions of normality (we used a Shapiro-Wilk test for normality on the residuals) and homogeneity of Variances (we used Levene's test for equality of variances) were not met for all waves.

TABLE 2 Machine learning US-CDC COVID-19 forecasting models- their description, features employed, methodology, and assumptions they make regarding public health interventions.

Model	Features used for forecasting	Author	Method	Assumptions
QJHong-Encounter (19)	Uses (1) Reproductive Number (R) and Encounter Density (D) relation in the past as a training set, (2) future D as input, and (3) ML/regression, the model predicts future R, and ultimately future daily new cases.	QJHong	Machine learning	Assumes current interventions will not change during the forecasted period.
OneQuietNight-ML (20)	Uses high-level features of daily case reports and movement trends data to make predictions about future Covid-19 cases.	OneQuietNight		
JHU_CSSE-DECOM (21)	County-level, empirical model driven by mobility, epidemiological, demographic, and behavioral data.	JHU CSSE		
UpstateSU-GRU (22)	A feed-forward RNN is used. The Seq2Seq algorithm trains the model to convert sequences from input to those in the output. The model inputs daily smoothed incident cases, deaths count, google mobility index, daily reproduction number, county demographic and health risk indices to model the baseline risk score.	SUNY Upstate and SU COVID-19 Prediction Team	County-level forecast using RNN seq2seq model.	

2.6 Statistical analysis

The statistical significance of all figures was determined by the non-parametric Kruskall-Wallis test which is followed by a Mann-Whitney test for groups that had a significant difference. Bar plots having a p-value <0.05 (statistically significant), are joined using an asterisk. In Figure 3, box plots are made representing the MAPE over all the predictions of a certain model for the corresponding forecast horizon. A box plot displays the distribution of data based on a five-number summary [“minimum”, first quartile (Q1), median, third quartile (Q3), and “maximum”]. We used the boxplot function (of `seaborn` library) in Python to plot it. `Seaborn` is a Python data visualization library based on `Matplotlib`.

3 Results

When case counts predicted by the various US CDC COVID-19 case prediction models are overlaid real-world data, visually several features stand out as illustrated in Figure 1A. On aggregate, models tend to approach the correct peak during various waves of the pandemic. However, some models undershot, others overshot, and many lagged the leading edge of real-world data by several weeks.

3.1 Complete timeline analysis

The MAPE values of all US CDC models were analyzed over the complete timeline and compared to two “Baselines”, which represented either an assumption that case counts would remain the same as the previous week (Baseline-I) or a simple linear model following the previous week’s case counts (Baseline-II) (see Figure 1B). Here, “IQVIA_ACOE-STAN,” “USACE-ERDC_SEIR,” “MSRA-DeepST,” and “USC-SI_kJalpha_RF” achieve the best performance with low MAPE ranging from 5 to 35%. In a comparison of overall performance, ensemble models performed significantly better than all other model types. However, the performance of ensemble models was not “significantly” better than the baseline models (no change or simple linear model) which performed better than both machine learning and epidemiological models overall (see Figure 1C). The peak predictions of all US-CDC models were plotted on the complete timeline and included in the study (see Figures 5–8).

3.2 Wave-wise analysis of best performing models

The MAPE values of all US CDC models were also analyzed wave-wise (see Figure 2). During the first wave of the pandemic, “Columbia_UNC-SurvCon" achieved the lowest MAPE = 14%, closely followed by “USACE-ERDC_SEIR" (MAPE = 17%) and “CovidAnalytics-DELPHI" (MAPE = 25%). Here, only 4 models performed better than both baselines. Three of these were epidemiological models while one was a hybrid model. From the bar plot of MAPE values of models categorized based on model type (refer Figure 1C), it can be inferred that during the first wave, hybrid models performed the best and attained the lowest MAPE. This was followed by the epidemiological models and those based on machine learning. On the other hand, ensemble models had the largest MAPE during this wave and none of them surpassed the Baseline-I MAPE, i.e., 31%.

During the second wave, “IQVIA_ACOE-STAN" performed the best with a MAPE score of 5% (see Figure 2). In this wave, a total of 13 models performed better than both baselines, with MAPE ranging from 5 to 37. These included five ensemble models, four epidemiological models, two machine learning models, and two hybrid models. All ensemble models exceed Baseline-I performance (that had MAPE = 37%), with the exception of “UVA-Ensemble". The epidemiological models showed a staggered MAPE distribution. Followed by the hybrid and the models categorized as “other" model sub-types, these have the lowest average MAPE in wave-II. In contrast to wave-I, ensemble models provide the best forecasts in wave-II (see Figure 1C). Here, hybrid models are the worst-performing models.

During the third wave (see Figure 1C), ensemble models performed similarly to wave-I. Baseline models had a relatively elevated high MAPE with Baseline-I and II MAPE scores being 74 and 77%, respectively. In wave-III, “USC-SI_kJalpha” is the best-performed model with MAPE= 32% (see Figure 2). Here, 32 models performed better than both baseline models. These included 12

TABLE 3 Ensemble learning US-CDC COVID-19 forecasting models- their description, features employed, methodology, and assumptions they make regarding public health interventions.

Model	Features used for forecasting	Author	Method	Assumptions
USC-SI_kJalpha (23)	Examines the impact of parameters learned via rapid linear regressions, emphasizing the reduction of hardware demands and achieving faster predictions. Using logistic regression, the model adjusts to samples for each variation within a specific day, considering dynamic patterns through a focus on recent data. Uses Random Forest to capture empirical errors in quantile projections, accounting for future trend changes.	USC	SIR	Current interventions will not change during forecasting period.
COVIDhub-ensemble (24)	An ensemble, or model average, of submitted forecasts to the COVID-19 Forecast Hub.	COVID-19 Forecast Hub	–	–
UCF-AEM (25)	Combines a traditional SEIR model with mixture modeling and uses ensemble neural networks to extract information from a complicated mixture modeling system.	UCF	SEIR model with ensemble neural networks.	Current interventions will not change during the forecasted period.
LNQ-ens1 (26)	Uses an ensemble of three models; two fit with LightGBM, and the third is a neural net. Ensemble weights are chosen each week manually based on performance in the previous week.	LockNQuay	Ensemble of three models.	Intervention effects are reflected in observable data and will continue in the future.
COVIDhub-baseline (27)	Baseline model for predictions. The most recent observed incidence is the median projection for all future horizons. From 1 week to next, the slope of the predicted medians for cumulative will be constant and equal to previously observed slope. The model looks at week-to-week incidence variations to generate a median distribution.	COVID-19 Forecast Hub	–	–
COVIDhub-trained_ensemble (28)	A weighted ensemble combination of all component model forecasts.	COVID-19 Forecast Hub	Ensemble	–
UVA-Ensemble (29)	Combines models using Bayesian model averaging. Auto-regressive method with features including mobility, other county case counts time-series, an LSTM model with mobility data as an additional predictor, and PatchSim, an SEIR variant with interaction between counties modeled using commuter data and calibrated on new confirmed cases.	University of Virginia, Bio-complexity COVID-19 Response Team	Ensemble of 3 models.	Impact of interventions is represented in observed data in 2 of 3 models, while the third assumes that interventions will change in the future.
Caltech CS156	Ensemble of 14 ML models: (a) Feedforward Neural Network, (b) Quantile Neural Network, (c) LSTM, (d) Conditional LSTM, (e) Encoder-Decoder Conditional LSTM, (f) Autoregressive, (g) Sessional Autoregressive, (h) Decision Tree, (i) Gradient-Boosted Decision Tree, (j) K-NN, (k) Gaussian Process, (l) Bayesian epidemiological, (m) Two-group epidemiological, (n) Curve-fitting.	Caltech	Ensemble of 14 models	–
MIT-Cassandra (30)	Based on the ensemble of predictions from four models, including (1) MDP feature representation, (2) KNN time-series, (3) Bi-LSTM time-series, (4) C-SEIRD epidemiological.	MIT Cassandra	Ensemble of four models	Assumes that current interventions will remain in place indefinitely.
FDANIHASU-Sweight (31)	Ensemble of submitted forecasts to COVID-19 Forecast Hub. The ensembles are formed by weighting the individual model forecasts with their past performances	FDANIHASU	Ensemble	–

compartment models, three machine learning models, four hybrid models, eight ensemble models, and five un-categorized models.

In the fourth wave of the pandemic, several models performed similarly between a MAPE of 28% and a baseline of 47% (Figure 2). Ensemble models performed the best whereas epidemiological models had the highest MAPE during this wave. Baseline-I and II MAPE scores were 47 and 48%, respectively. In wave-IV, “LANL-GrowthRate” is the best performed model with MAPE = 28% (Figure 2). In the fourth wave, 17 models performed better than both baseline models. These included six compartment models, seven ensemble models, one machine learning models, two hybrid models, and one uncategorized models.

3.3 Effect of increase in forecast horizon

The MAPE of models in the US CDC database increased each week out from the time of prediction. Figure 3 depicts a strictly increasing rise in MAPE with the increase in the forecast horizon. In other words, the accuracy of predictions declined the further out they were made. At 1 week from the time of prediction, the MAPE of models examined clustered just below 25% MAPE and declined to about 50% MAPE by 4 weeks.

The MAPE in each week was relatively bi-modal, with several models fitting within a roughly normal distribution and others having a higher MAPE. The distribution of the points indicates that the majority of models project similar predictions for smaller forecast horizons, while the predictions for larger horizons are more spread out. The utility of model interpretation would improve by excluding those models that fall more than one standard deviation (σ) from the average MAPE of models.

Though we examined 54 models overall (reported 51 in the full timeline), many of these did not make predictions during the first wave of the pandemic when data was less available and therefore do not appear in the wave-wise MAPE calculation and subsequent analysis. The number of weeks for which each model provides predictions was highly variable, as seen in Figure 4.

TABLE 4 Hybrid US-CDC COVID-19 forecasting models- their description, features employed, methodology, and assumptions they make regarding public health interventions.

Model	Features used for forecasting	Author	Method	Assumptions
JHUAPL-Bucky (32)	Uses public mobility data to build a Spatial compartment model.	JHUApplied Physics Lab	Spatial compartment model	—
FRBSF_Wilson-Econometric (33)	Econometric model that connects the current transmission rate with the fraction of the population that is vulnerable to the shift in new infections from now until a future horizon. Current transmission rate is assumed to be caused by people's mobility and the weather. Mobility, weather, and acquired natural immunity are accounted for; Includes infection growth lag, implying infection growth lag predicts future infection growth. County-specific intercepts are introduced, which allows each county to have a distinct mean level of infection increase.	Federal Reserve Bank of San Francisco/Wilson	SIR-derived econometric county panel data model	Intervention effects are reflected in observable data and will continue in future.
IEM_MED-CovidProject (34)	Uses an AI model to fit data from various sources and project new cases of COVID-19. Assumes that the *R*-value (average number of secondary infections) changes quite rapidly over time due to changes in human behavior and uses a sliding window that fits the data and finds the best *R*-values for each window.	IEM MED	SEIR model with ML	Current interventions will not change during the forecasted period.
MOBS-GLEAM_COVID (35)	A metapopulation method is used. The world is divided into geographical subpopulations, and human mobility between subpopulations is depicted on a network. This data layer on mobility identifies the number of persons traveling from between sub-populations. The mobility network is made up of many mobility processes, ranging from short-distance commuting to intercontinental travel. Superimposed on the globe population and mobility layers is an agent-based epidemic model that describes the infection and population dynamics.	MOBS Lab at Northeastern	Metapopulation age-structured SLIR	Social distancing policies in place at the date of calibration are extended for the future weeks.
DDS-NBDS (36)	Jointly modeling daily deaths and cases using a negative binomial distribution based non-parametric Bayesian generalized linear dynamical system (NBDS).	Team DDS	Bayesian hierarchical model	Intervention effects are reflected in observable data will continue in future.

4 Discussion

Accurate modeling is critical in pandemics for a variety of reasons. Policy decisions need to be made by political entities that must follow procedures, sometimes requiring weeks for a proposed policy intervention to become law and still longer to be implemented. Likewise, public health entities such as hospitals, nurseries, and health centers need "lead time" to distribute resources such as staffing, beds, ventilators, and oxygen supplies. However, modeling is limited, especially by the availability of data, particularly in early outbreaks (46). Resources such as ventilators are often distributed heterogeneously (47), leading to a risk of unnecessary mortality. Similarly, the complete homogeneous distribution of resources like masks is generally sub-optimal and may also result in deaths (48). Likewise, it is important to develop means of assessing which modeling tools are most effective and trustworthy. For example, a case forecasting model that consistently makes predictions that fare worse than assuming that case counts will remain unchanged or that they will follow a simple linear model isn't likely to be useful in situations where modeling is critical. The use of these baselines allows for the exclusion of models that "fail" to predict case counts adequately.

Direct comparison of error based on the difference from real-world data potentially excludes important dimensions of model accuracy. For example, a model that accurately predicts the time course of disease cases (but underestimates cases by 20% at any given point) might have greater utility for making predictions about when precautions are necessary when compared to a model that predicts case numbers with only a 5% error but estimates peak cases 2 weeks late. To assess the degree of timing error, the peak of each model was compared to the true peak of cases that occurred within the model time window. MAPE i.e., the ratio of the error between the true case count and a model's prediction to the true case count, is a straightforward way of representing the quality of predictive models and comparing between model types. MAPE has several advantages over other metrics of measuring forecast error. First, since MAPE deals with negative residuals by taking an absolute value rather than a squared one, reported errors are proportional. Second, since it's a percent error, it is also conceptually straightforward to understand. Using absolute error as a measure might be misleading at times. For example, a model that predicts 201, 000 cases when 200, 000 cases occur has good accuracy but the same absolute error as a model that predicts 1, 100 cases when 100 occur. MAPE accounts for this by normalizing by the number of data points.

The average MAPE of successful models which we define as lower MAPE than either baseline model varied between methods depending on the wave they were measured in. In the first wave, epidemiological models had a mean MAPE of 31%, and machine learning models had a mean MAPE of 32%. In the second wave, these were 45 and 44%, respectively. In the third wave, these were 63 and 63%, and in the fourth wave, these were 92 and 44%, respectively. Therefore, we notice that the mean MAPE of models got worse with each wave. This is because each model type is susceptible to changing real-world conditions, such as the emergence of new variants with the potential to escape prior immunity, or a higher R_0, new masking or lockdown mandates, the spread of conspiracy theories, or the development of vaccines which will decrease the number of individuals susceptible to infection.

In waves-2 and 3, the best performing model was a hybrid and machine learning model, respectively. In waves, 1 and 4, the best model was an epidemiological and an "other" model,

TABLE 5 Other US-CDC COVID-19 forecasting models- their description, features employed, methodology and assumptions they make regarding public health interventions.

Model	Features used for forecasting	Author	Method	Assumptions
IBF-TimeSeries (37)	Combines mechanistic disease transmission model with a curve-fitting approach.	Institute of Business Forecasting	Modified Time Series Model	Do not make any specific assumptions.
RobertWalraven (38)	Uses a skewed Gaussian distribution with four empirical parameters: height, position, left growth rate, and right decay rate. The model makes no epidemiological assumptions and has no epidemiological parameters.	Robert Walraven	Skewed Gaussian distribution	Current interventions will not change during the forecasted period.
UMich-RidgeTfReg (39)	This model is based on ridge regression (penalized Ordinary Least Squares regression) to make predictions without relying on external assumptions. The model uses Finite Impulse Response filtering to forecast confirmed cases each day as a function of prior day numbers.	UMich	Ridge regression	Current interventions will not change during the forecasted period.
Karlen-pypm (40)	Uses Discrete-time difference equations with long periods of the constant transmission rate.	Karlen Working Group	Discrete-time difference equations	Intervention effects are reflected in observable data and will continue in the future.
LANL-GrowthRate (41)	Presents two processes: first statistical model depicts how the number of COVID-19 infections varies over time while the second model correlates the number of infections with the reported data. The underlying numbers of susceptible and infected cases in the population at the preceding time step, scaled by the size of the state's initial susceptible population, are used to map the rise of new instances. These two components' weights are dynamically adjusted.	Los Alamos National Labs	Statistical dynamical growth model	Interventions on the first day of the forecast will continue over the following 4 weeks.
JCB-PRM (42)	Built on observations of macro-level societal and political responses to COVID-19 characterized only in terms of infections and deaths. Assumes that the actual net impact of policy actions undertaken, though not identical across time or geography, is predictable. Identifies acceptability ranges from observation data up to current time.	John Burant	Phenomeno-logical statistical model	Incidence of COVID-19 in the population determines the strength and impact of future control measures.
SigSci-TS (43)	Time series forecasting using ARIMA for case forecasts and lagged cases for death forecasts.	Signature Science FOCUS	Autoregressive time-series model	Current interventions will not change during the forecasted period.
CEID-Walk (44)	The model is based on a random walk with no drift. The variance in step size of random walk is estimated using the last few observations of a target time series.	University of Georgia CEID Forecasting Working Group	Statistical random walk model	Social distancing policies in place at the date of calibration are extended for the future weeks.
MIT_ISOLAT-Mixtures (45)	A non-mechanistic, non-parametric forecasting model that forecasts time series as a sum of bell curves. The confidence intervals are calculated by applying a multiplicative log-Gaussian perturbation to the observed time series.	IDSS COVID-19 Collab. at MIT Mixture model	—	Current interventions will remain in place indefinitely.

respectively (see Figure 2). In each wave, some examples of each model type were successful, and some were unsuccessful. After wave 1, ensemble and ML models on average had the lowest MAPE but no model category significantly outperformed baseline models (Figure 1C). These results suggest that no overall modeling technique is inherently superior for predicting future case counts.

Compartmental (or epidemiological) models broadly use several "compartments" which individuals can move between such as "susceptible," "infectious," or "recovered," and use real-world data to arrive at estimates for the transition rate between these compartments. However, the accuracy of a compartment model depends heavily on accurate estimates of the R_0 in a population, a variable that changes over time, especially as new virus variants emerge. For example, the emergence of the Iota variant of SARS-CoV-2 (also known as lineage B.1.526) resulted in an unpredicted increase in case counts (49). Machine learning models train algorithms that would be difficult to develop by conventional means. These models "train" on real-world data sets and then make predictions based on that past data. Machine learning models are sensitive to the datasets they are trained on, and small or incomplete datasets produce unexpected results. Hybrid models make use of both compartment modeling and machine learning tools. On the other hand, ensemble models combine the results of multiple other models hoping that whatever errors exist in the other models will "average out" of the combined model. Ensemble models potentially offer an advantage over individual models in that by averaging the predictions of multiple models, flawed assumptions or errors in individual models may "average out" and result in a more accurate model. However, if multiple models share flawed assumptions or data, then averaging these models may simply compound these errors. An individual model may achieve a lower error than multiple flawed models in these cases.

The "baseline" models had a MAPE of 48 and 54% over the entire course of the pandemic in the US. Most models did not perform better than these. Among those that did, we discuss the five with the best performance (in increasing order of their performance). First, "QJHong-Encounter" is a model by Qijung Hong, an Assistant professor at Arizona State University. This model uses an estimate of encounter density (how many potentially infectious encounters people are likely to have in a day) to predict changes in estimated R (reproduction number), and then uses that to estimate future daily new cases. The model uses machine learning. It had a MAPE of 38% over all waves. Second, "USC-SI_kJalpha_RF" is a hybrid model from the University of Southern

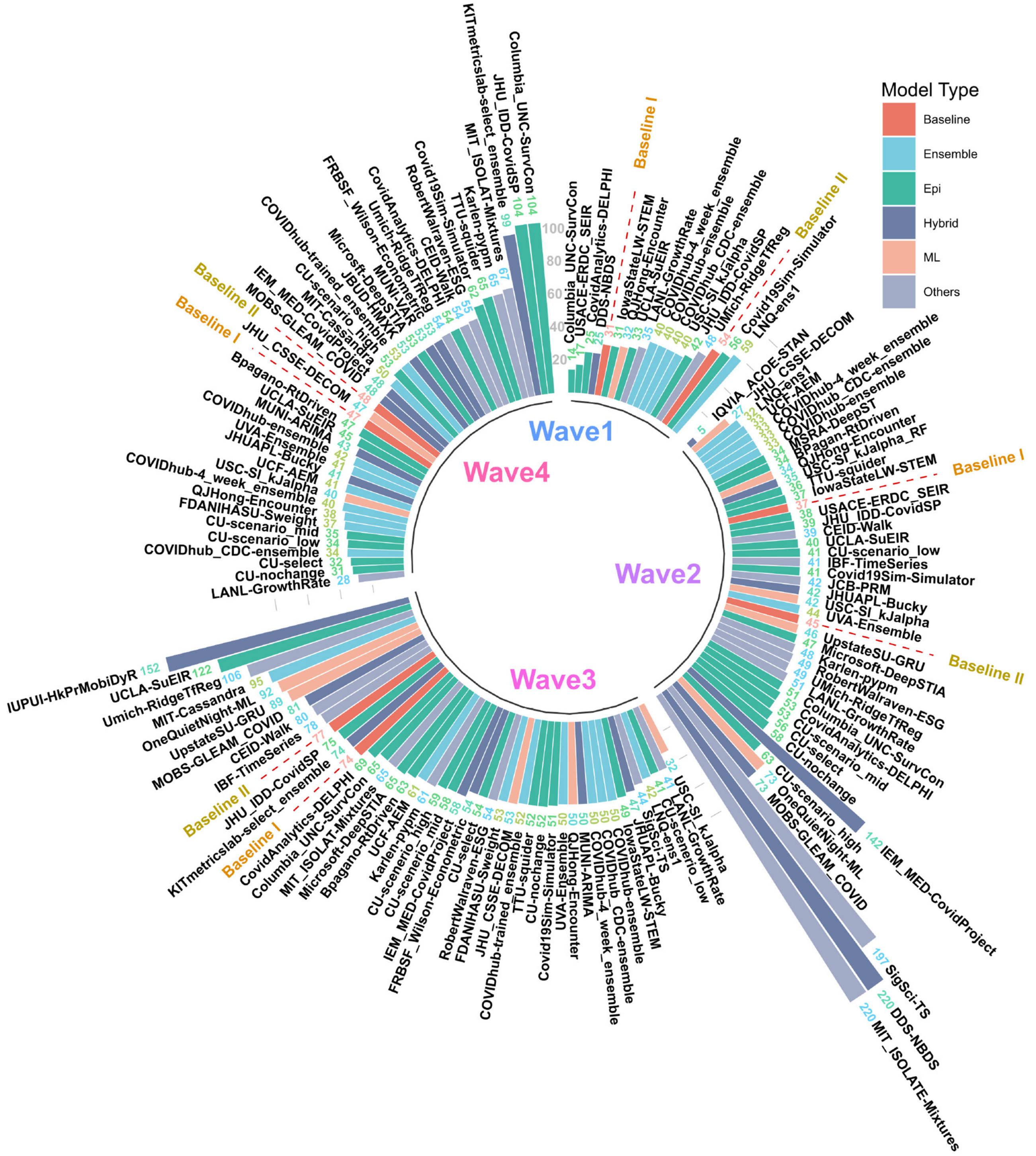

FIGURE 2
MAPE values of US CDC Forecasting models in wave-I to IV. Models are sorted in descending order of MAPE. The color scheme represents the model category. Here "Baselines" are represented in red.

California Data Science Lab. This model also uses a kind of hybrid approach, where additional parameters are modeled regionally for how different regions have reduced encounters and machine learning is used to estimate parameters (23). It had a MAPE of 35% over all waves. Third, "MSRA-DeepST" is a SIR hybrid model from Microsoft Research Lab-Asia that combines elements of SEIR models and machine learning. It had a MAPE of 34%. Next, "USACE-ERDC_SEIR" is a compartment model developed by the US Army Engineer Research and Development Center COVID-19 Modeling and Analysis Team. It adds to the classic SEIR model,

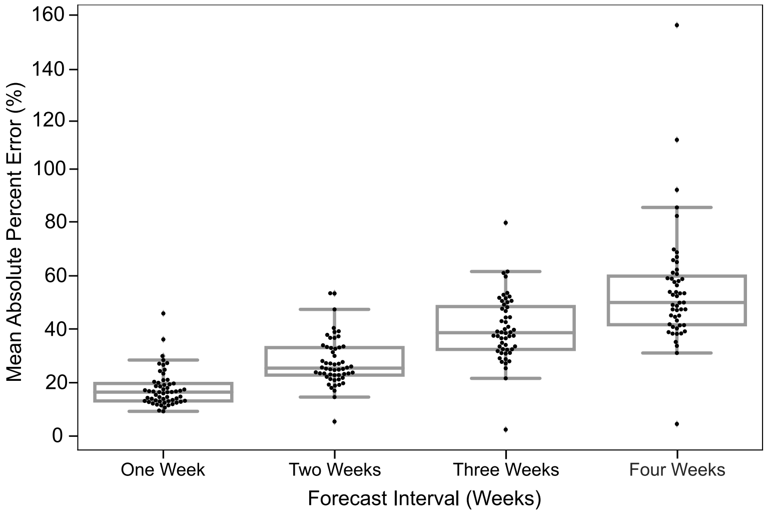

FIGURE 3
Predictions are most accurate closest to the time of prediction. The MAPE in predictions of all models for different forecast horizons is shown. The dots in each box plot represent the MAPE over all the predictions of a certain model for the corresponding forecast horizon. The *y*-axis is the MAE between the predicted case count and the reported case count. The *x*-axis is the forecast horizon.

adding additional compartments for unreported infections and isolated individuals. It used Bayesian estimates of prior probability based on subject matter experts to select initial parameters. It had a MAPE of 31%. Lastly, "IQVIA_ACOE_STAN" is a machine learning model from IQVIA-Analytics Center of Excellence and has the highest apparent performance on the overall timeframe. The calculated MAPE for this model was 5%. Notably, the MAPE of 5% is much lower than the MAPE of 31% of the next closest model.

Although MAPE is superior to other methods of comparing models, there are still some challenges. For example, "IQVIA_ACOE_STAN" appears to be the lowest model by far, but the only data available covers a relatively short time frame, and unfortunately discontinued its contribution to the CDC website after Wave 2. The time frame that it predicted also does not include any changes in case direction from upswings and downswings. This advantages the model compared to other models, which might cover time periods where case numbers peak or new variants emerge. This highlights a potential pitfall of examining this data: the models are not studying a uniform time window.

Data reporting is one of the sources of error affecting model accuracy. There is significant heterogeneity in the reporting of COVID-19 cases by state. This is caused by varying state laws, resources made available for testing, the degree of sequencing being done in each state, and other factors. Additionally, different states have heterogeneity in vaccination rate, population density, implementation of masking and lockdown, and other measures that may affect case count predictions. Therefore, the assumptions underlying different models and the degree to which this heterogeneity is taken into account may result in models having heterogeneous predictive power in different states. Although this same thinking could be extended to the county level, the case count reporting in each county is even more variable and makes comparisons difficult.

To make the comparison between models more even, we used multiple times segments to represent the various waves during the pandemic. Models that have made fewer predictions, particularly avoiding the "regions of interest" such as a fresh wave or a peak, would only be subjected to a less challenging evaluation than the models that covered most of the timeline. Further, the utility of models with only a few predictions within "regions of interest" is also questionable. Considering these aspects, we define four time segments corresponding to each of the major waves in the US. Within each of these waves of interest, we consider only models that made a significant number of predictions for the purpose of comparison. Such a compartmentalized comparison is now straightforward, as all models within a time segment can now have a common evaluation metric.

We focus on evaluating the model performance for their 4-week ahead forecasts. A larger forecast horizon provides a higher real-world utility in terms of policy-making or taking precautionary steps. We believe this to be a more accurate representation of a model's predictive abilities as opposed to smaller windows in which desirable results could be achieved by simply extrapolating the present trend. Therefore, due to the time delays associated with policy decisions and the movement of critical resources and people, the long-term accuracy of models is of critical importance. To take an extreme example, a forecasting model that only predicted 1 day in advance would have less utility than one that predicted ten days in advance.

The failure of roughly one-third of models in the CDC database to produce results superior to a simple linear model should raise concerns about hosting these models in a public venue. Without strict exclusion criteria, the public may not be aware that the are significant differences in the overall quality of these models. Each model type is subject to inherent weaknesses of the available data. The accuracy of compartment models is heavily dependent on the quality and quantity of reported data and also depends on a

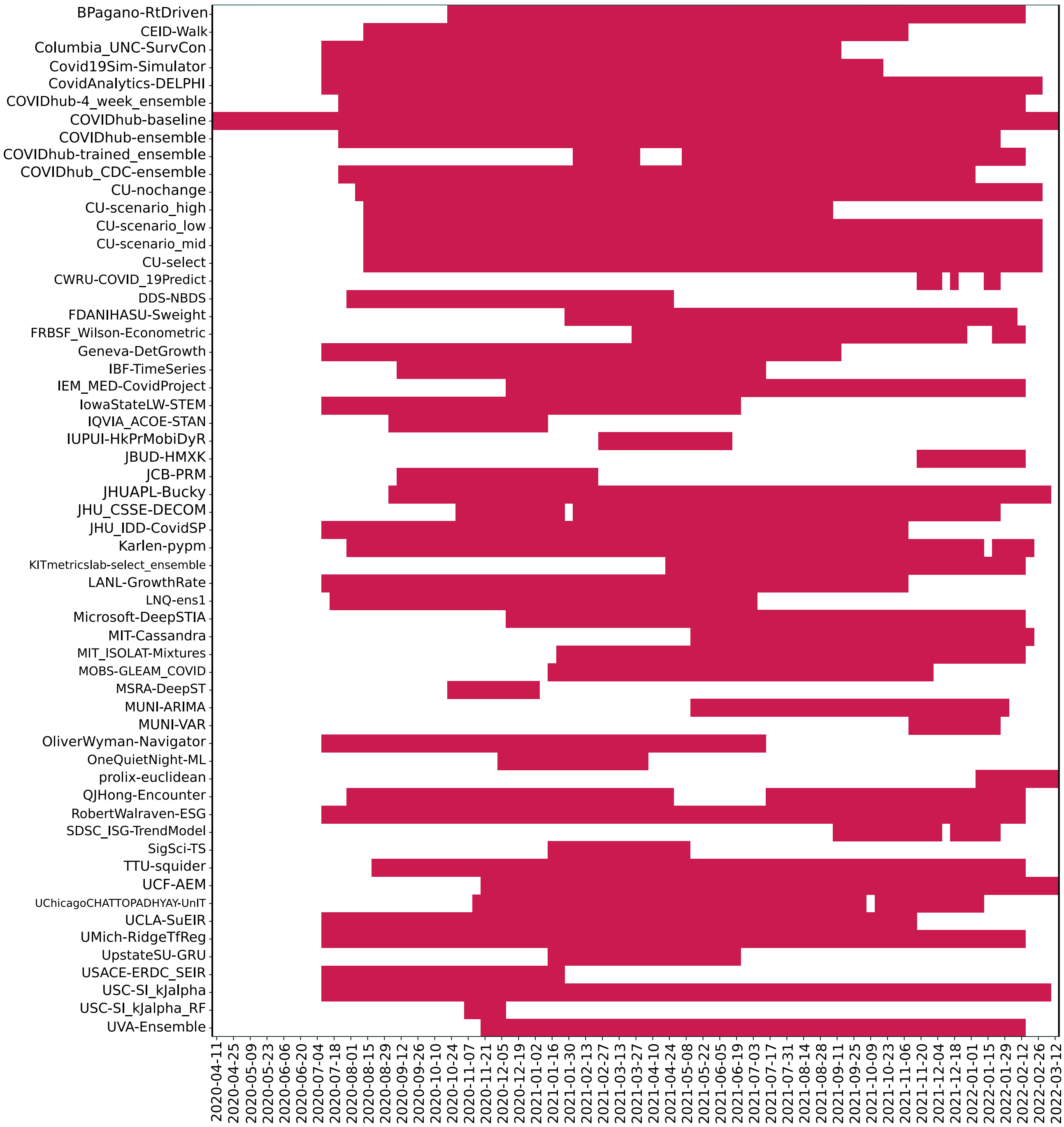

FIGURE 4
Plot depicting frequency of 04 week ahead predictions made by models. Here, models were ordered alphabetically on the *y*-axis. The *x*-axis represents the target dates for which the predictions were made. Dates range from July 2020 to Jan 2022.

variable that might change with the emergence of new variants. Heterogeneous reporting of case counts, variable accuracy between states, and variable early access to testing resulted in limited data sets. Likewise, it seems that since training sets did not exist, machine learning models were unable to predict the Delta variant surge. Robust evidence-based exclusion criteria and performance-based weighting have the potential to improve the overall utility of future model aggregates and ensemble models.

Because the US CDC has a primary mission focused on the United States, the models included are focused on United States case counts. However, globally the assumptions necessary to produce an accurate model might differ due to differences in population density, vaccine availability, and even cultural beliefs about health. However, identifying the modeling approaches that work best in the United States provides a strong starting point for global modeling. Some of the differences in modeling will be

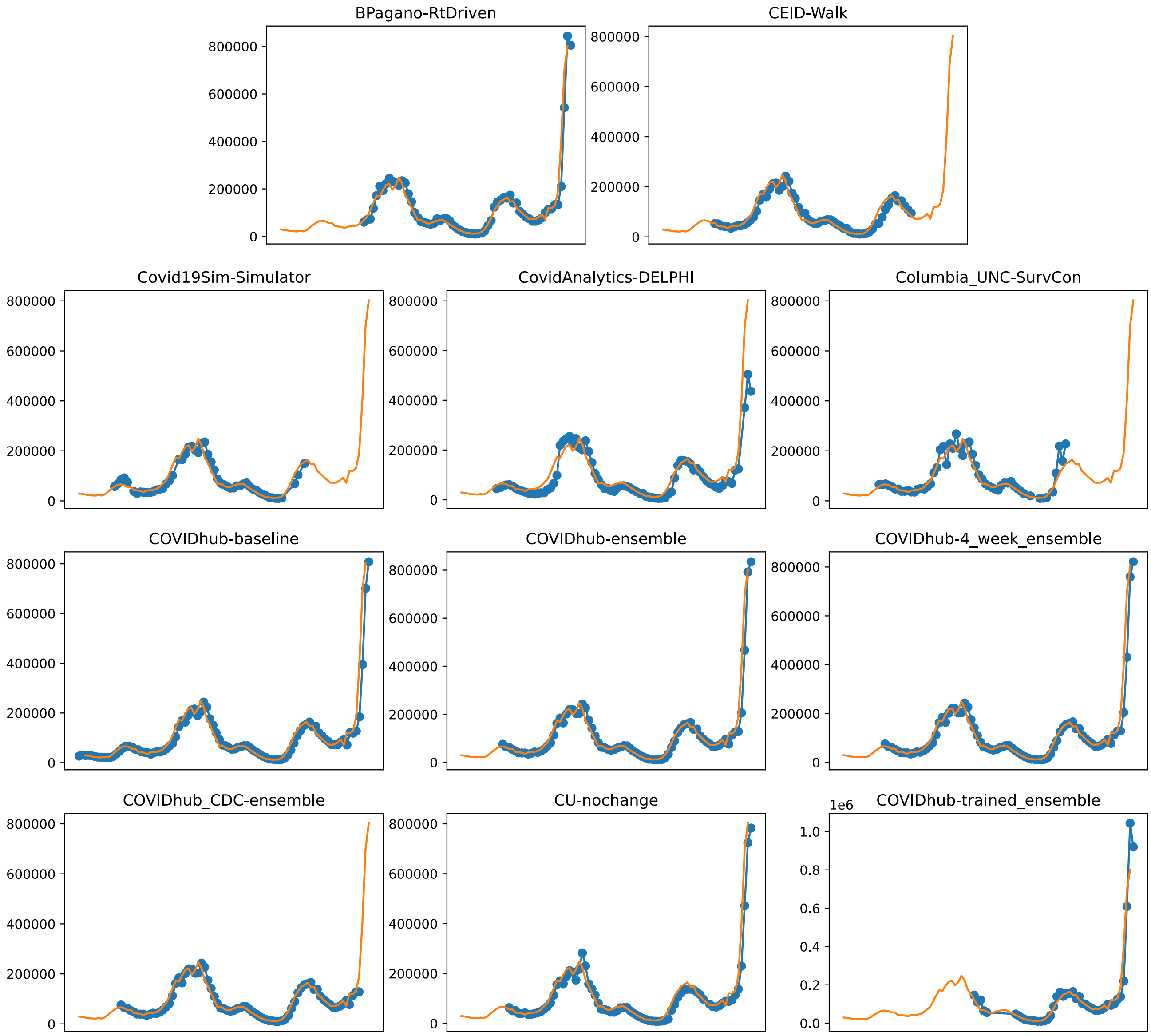

FIGURE 5
Plots depicting the peak predictions of US-CDC models—"BPagano-RtDriven," "CEID-Walk," "Covid19Sim-Simulator," "CovidAnalytics-DELPHI," "Columbia_UNC-SurvCon," "COVIDhub-baseline," "COVIDhub-ensemble," "COVIDhub-4_week_ensemble," "COVIDhub_CDC-ensemble," "CU-nochange," "COVIDhub-trained_ensemble," over complete timeline.

accounted for by different input data, which can be customized by country or different training sets in the case of machine learning models.

The ultimate measure of forecasting model quality is whether the model makes a prediction that is used fruitfully to make a real-world decision. Staffing decisions for hospitals can require a lead time of 2–4 weeks to prevent over-reliance on temporary workers, or shortages (50). Oxygen has become a scarce resource during the COVID-19 pandemic and also needs lead time (51). This has had real-world policy consequences as public officials have ordered oxygen imports well after they were needed to prevent shortages (52). Indeed for sufficient time to be available for public officials to enact new policies and for resources to be moved, a time frame of 8 weeks is preferable. The need for accurate predictions weeks in advance is confounded by the declining accuracy of models multiple weeks in advance, especially considering the rise time of new variant waves (Figure 3). During the most recent wave of infections, news media reported on the potential for the "Omicron" variant of SARS-CoV-2 to rapidly spread in November of 2021, however in the United States, an exponential rise was not apparent in case counts until December 14th, when daily new case counts approximated 100 k new cases per day, and by January 14th, 2022, new cases exceed 850, 000 new cases per day. The time when accurate predictive models are most useful is ahead of rapid rises in cases, something none of the models examined were able to predict, given the rise in SARS-CoV-2 cases that occurred during the pandemic.

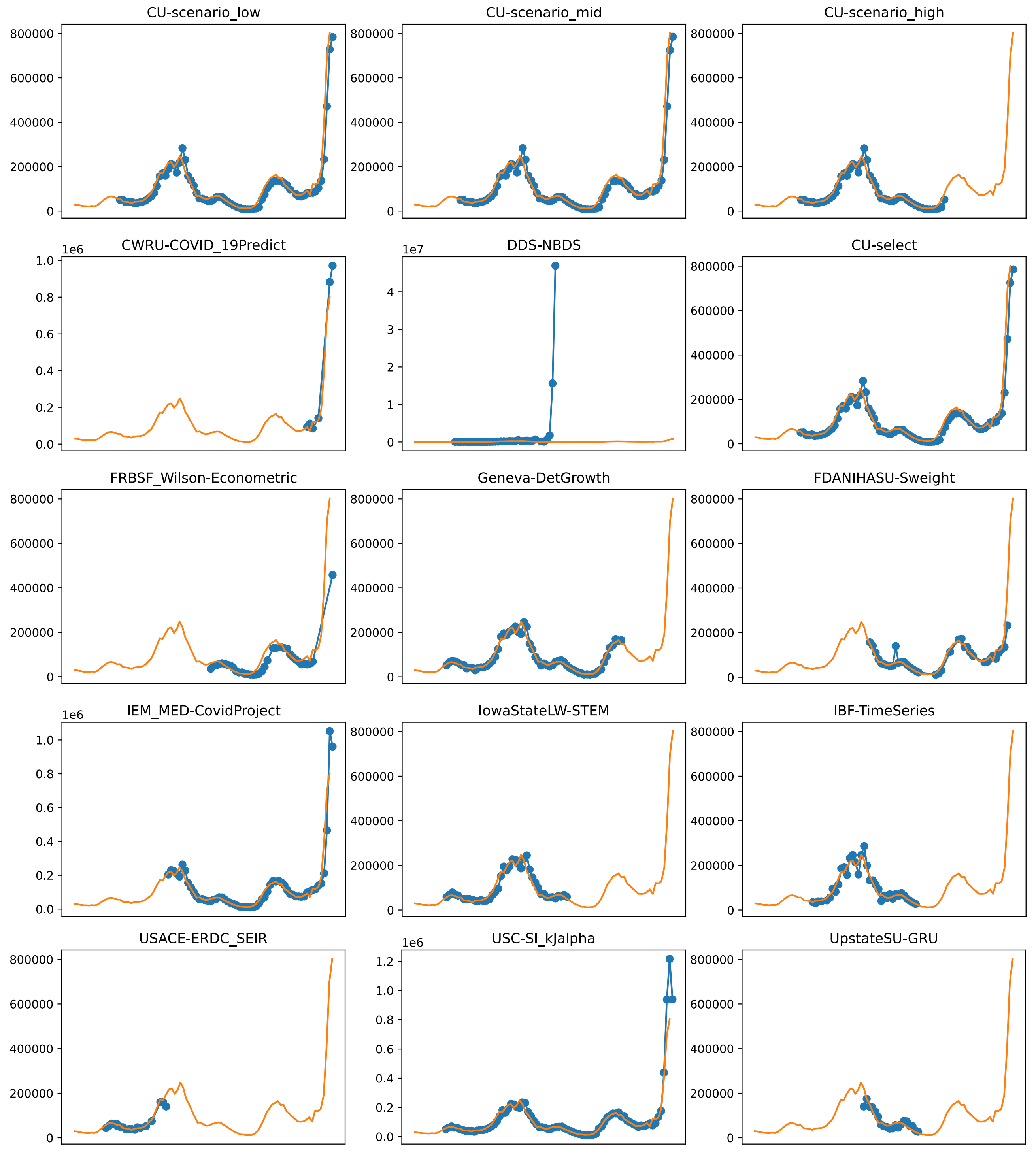

FIGURE 6
Plots depicting the peak predictions of various US-CDC models—"CU-scenerio_low," "CU-scenerio_mid," "CU-scenerio_high," "CWRU-COVID_19Predict," "DDS-NBDS," "CU-select," "FRBSF_Wilson-Econometric," "Geneva-DetGrowth," "FDANIHASU-Sweight," "IEM_MED-CovidProject," "IowaStateLW-STEM," "IBF-TimeSeries," "USACE-ERDC_SEIR," "USC-SI_KJalpha," "UpstateSU-GRU," over complete timeline.

5 Conclusion

Although forecasting models have gained immense attention recently, many challenges are faced in developing these to serve the needs of governments and organizations. We found that most models are not better than CDC baselines. The benchmark for resource allocation ahead of a wave still remains the identification and interpretation of new variants and strains by biologists. We propose a reassessment of the role of forecasting models in pandemic modeling. These models as currently

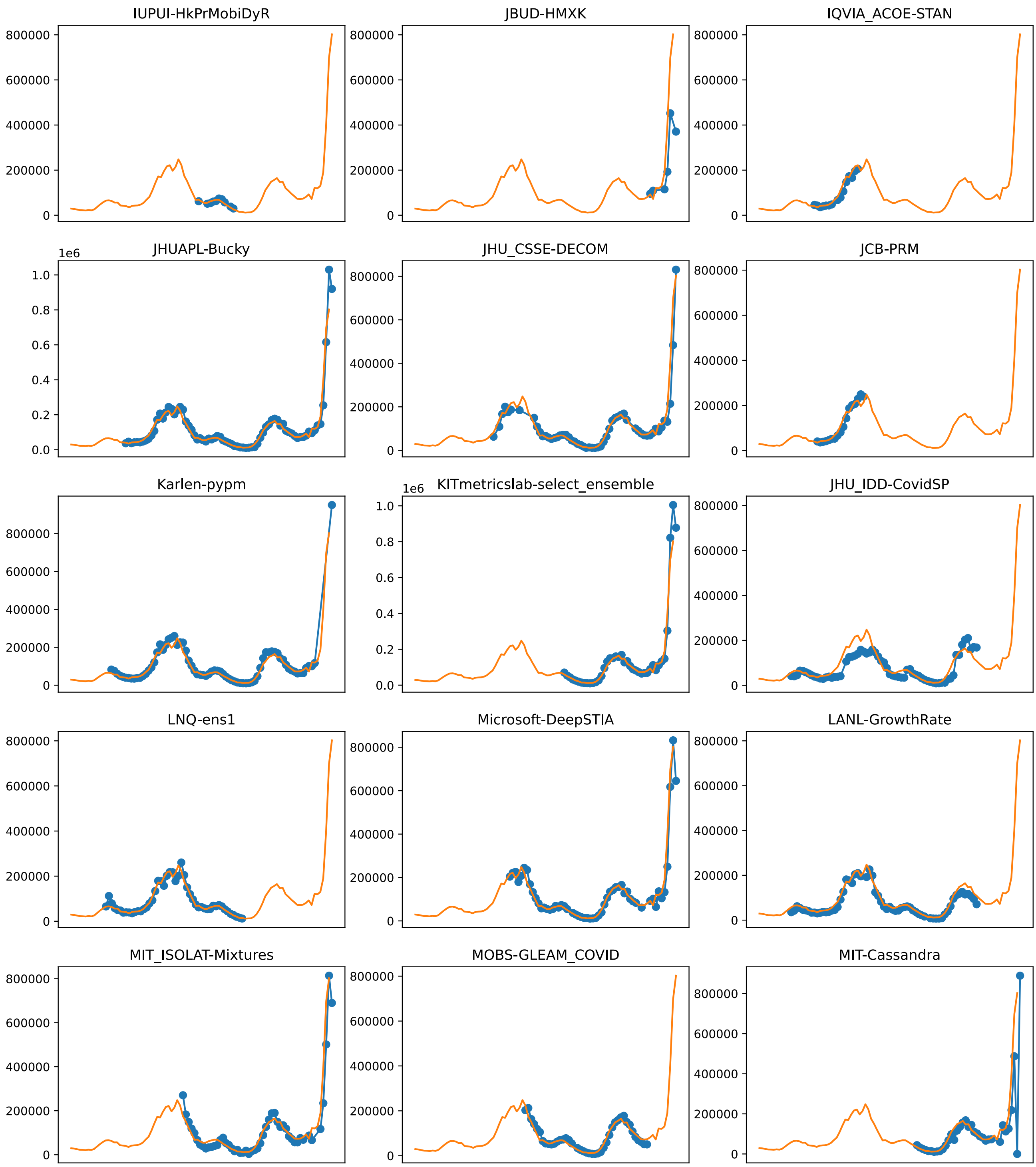

FIGURE 7
Plots depicting the peak predictions of various US-CDC models—"IUPUI-HkPrMobiDyR," "JBUD-HMXK," "IQVIA_ACOE-STAN," "JHUAPL-Bucky," "JHU_CSSE-DECOM," "JCB-PRM," "Karlen-pypm," "KITmetricslab-select_ensemble," "JHU_IDD-CovidSP," "LNQ-ens1," "Microsoft-DeepSTIA," "LANL-GrowthRate," "MIT_ISOLAT-Mixtures," "MOBS-GLEAM_COVID," "MIT-Cassandra," over complete timeline.

implemented can be used to predict the peak and decline of waves that have already been initiated and can provide value to decision-makers looking to allocate resources during an outbreak. Prediction of outbreaks beforehand however still requires "hands-on" identification of cases, sequencing, and data gathering. The lack of clean, structured, and accurate datasets also affects the performance of the case prediction models, making the estimation of patient mortality count and transmission rate significantly harder. More robust sources of data on true case numbers, variants, and immunity

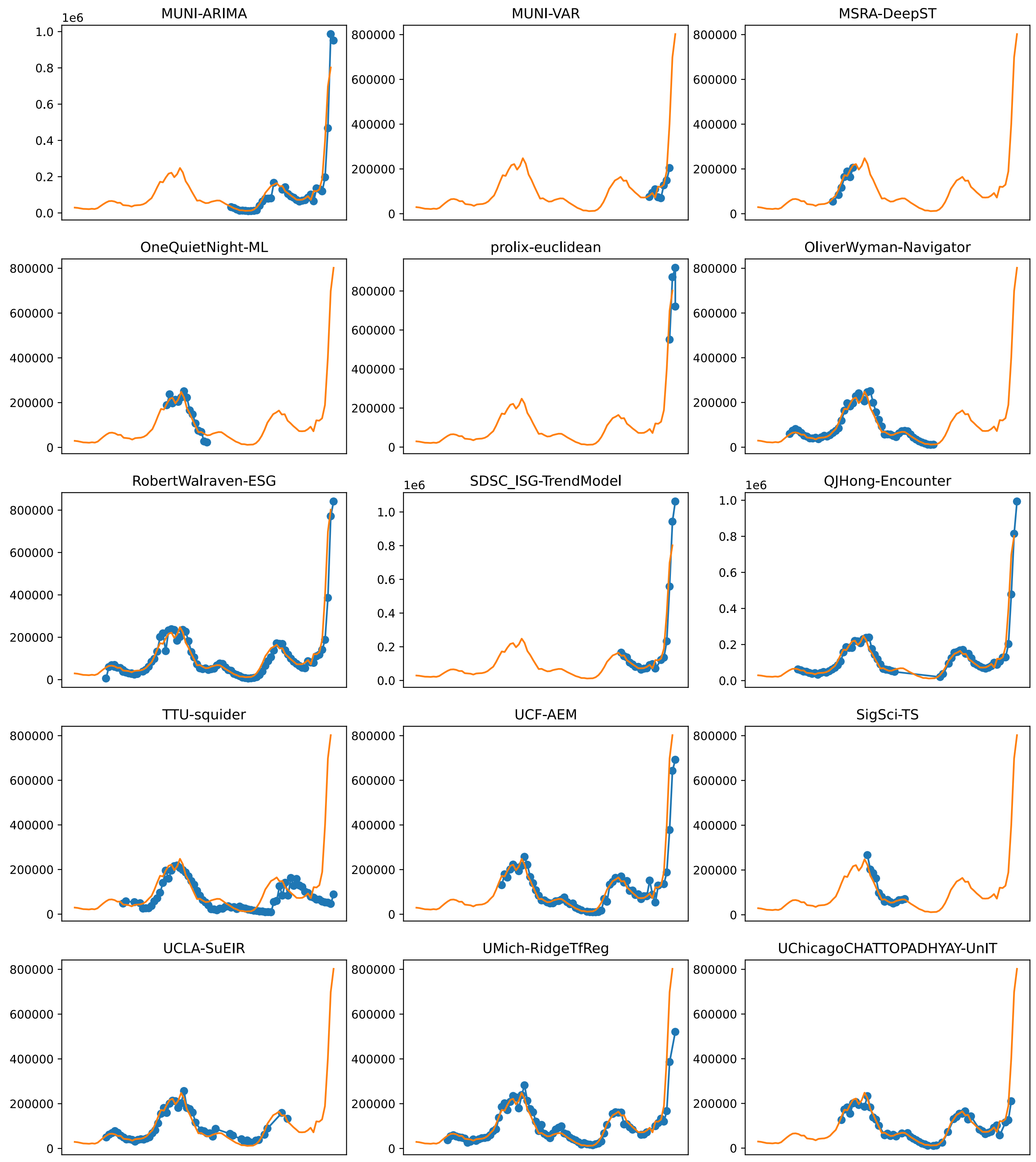

FIGURE 8
Plots depicting the peak predictions of various US-CDC models—“MUNI-ARIMA,” “MUNI-VAR,” “MSRA-DeepST,” “OneQuietNight-ML,” “prolix-euclidean,” “OliverWyman-Navigator,” “RobertWalraven-ESG,” “SDSC_ISG-TrendModel,” “QJHong-Encounter,” “TTU-squider,” “UCF-AEM,” “SigSci-TS,” “UCLA-SuEIR,” “UMich-RidgeTfReg” and “UChicagoCHATTOPADHYAY-UnIT” over complete timeline.

would be useful to create more accurate models that the public and policymakers can use to make decisions. Future work can focus on framing a novel evaluation metric for measuring the time lag and direction of prediction in the short term.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

AC: Writing – original draft, Writing – review & editing, Conceptualization, Data curation, Project administration. GJ: Writing – review & editing, Data curation, Formal analysis, Investigation, Methodology. RJ: Writing – review & editing, Visualization. ML: Visualization, Writing – review & editing. JB: Supervision, Validation, Writing – original draft, Writing – review & editing. CG: Conceptualization, Project administration, Resources, Supervision, Validation, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A novel Coronavirus from patients with pneumonia in China, 2019. *N Engl J Med.* (2020) 382:727–33. doi: 10.1056/NEJMoa2001017

2. 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. doi: 10.1016/S1473-3099(20)30120-1

3. CDC. *COVID Data Tracker*. (2020). Available online at: https://covid.cdc.gov/covid-data-tracker (accessed April 15, 2022).

4. Albani V, Loria J, Massad E, Zubelli J. COVID-19 underreporting and its impact on vaccination strategies. *BMC Infect Dis.* (2021) 21:1111. doi: 10.1186/s12879-021-06780-7

5. Stadlbauer D, Tan J, Jiang K, Hernandez MM, Fabre S, Amanat F, et al. Repeated cross-sectional sero-monitoring of SARS-CoV-2 in New York City. *Nature.* (2021) 590:146–50. doi: 10.1038/s41586-020-2912-6

6. Alakus TB, Turkoglu I. Comparison of deep learning approaches to predict COVID-19 infection. *Chaos Solit Fract.* (2020) 140:110120. doi: 10.1016/j.chaos.2020.110120

7. Fonseca I, Casas P, García i Carrasco V, Garcia i Subirana J. SEIRD COVID-19 formal characterization and model comparison validation. *Appl Sci.* (2020) 10:5162. doi: 10.3390/app10155162

8. Khan ZS, Van Bussel F, Hussain F. A predictive model for Covid-19 spread applied to eight US states. *Epidemiol Infect.* (2020) 148:e249. doi: 10.1017/S0950268820002423

9. Lemaitre JC, Grantz KH, Kaminsky J, Meredith HR, Truelove SA, Lauer SA, et al. A scenario modeling pipeline for COVID-19 emergency planning. *Sci Rep.* (2021) 11:7534. doi: 10.1038/s41598-021-86811-0

10. Wang L, Wang G, Gao L, Li X, Yu S, Kim M, et al. Spatiotemporal dynamics, nowcasting and forecasting of COVID-19 in the United States. *arXiv* [preprint]. (2020). doi: 10.1090/noti2263

11. Pagano B. *Covid-19 Modeling - bob pagano*. (2020). Available online at: https://bobpagano.com/covid-19-modeling/ (accessed April 15, 2022).

12. Zou D, Wang L, Xu P, Chen J, Zhang W, Gu Q. Epidemic model guided machine learning for COVID-19 forecasts in the United States. *medRXiv* [preprint]. (2020). doi: 10.1101/2020.05.24.20111989

13. Chhatwal J. *Covid19sim-Simulator*. (2020). Available online at: https://covid19sim.org/ (accessed April 15, 2022).

14. Mayo ML, Rowland MA, Parno MD, Detwiller ID, Farthing MW, England WP, et al. *USACE-ERDC_SEIR*. (2020). Available online at: https://github.com/erdc-cv19/seir-model (accessed April 15, 2022).

15. Gao Z, Li C, Cao W, Zheng S, Bian J, Xie X, et al. *Microsoft-DeepSTIA*. (2020). Available online at: https://www.microsoft.com/en-us/ai/ai-for-health (accessed April 15, 2022).

16. Li ML, Bouardi HT, Lami OS, Trichakis N, Trikalinos T, Zarandi MF, et al. *Overview of DELPHI model V3 - COVIDAnalytics*. (2020). Available online at: https://www.covidanalytics.io/DELPHI_documentation_pdf (accessed April 15, 2022).

17. Wang Q, Xie S, Wang Y, Zeng D. Survival-convolution models for predicting COVID-19 cases and assessing effects of mitigation strategies. *Front Public Health.* (2020) 8:325. doi: 10.3389/fpubh.2020.00325

18. Pei S, Shaman J. Initial simulation of SARS-CoV2 spread and intervention effects in the continental US. *medRxiv* [preprint]. (2020). doi: 10.1101/2020.03.21.20040303

19. Hong QJ. *QJHong-Encounter: Parameter-Free Covid-19 Model Based on Encounter Density Data*. (2020). Available online at: https://github.com/qjhong/covid19 (accessed April 15, 2022).

20. Jo A, Cho J. *OneQuietNight-ML Covid-19 Forecast*. (2020). Available online at: https://github.com/One-Quiet-Night/COVID-19-forecast (accessed April 15, 2022).

21. Marshall M, Gardner L, Drew C, Burman E, Nixon K. *JHU_CSSE-DECOM*. (2020). Available online at: https://systems.jhu.edu/research/public-health/predicting-covid-19-risk/ (accessed April 15, 2022).

22. Zhang-James Y, Salekin A, Hess J, Chen S, Wang D, Morley CP, et al. *UpstateSU-GRU*. (2021). Available online at: https://github.com/ylzhang29/UpstateSU-GRU-Covid (accessed April 15, 2022).

23. Srivastava A, Xu T, Prasanna VK. Fast and accurate forecasting of COVID-19 deaths using the SIkJα model. *arXiv* [preprint]. (2020). doi: 10.48550/arXiv.2007.05180

24. Wattanachit N, Ray EL, Reich N. *COVIDhub-Ensemble*. (2020). Available online at: https://github.com/reichlab/covid19-forecast-hub (accessed April 15, 2022).

25. Wang D, Summer T, Zhang S, Wang L. *UCF-AEM*. (2020). Available online at: https://github.com/UCF-AEM/UCF-AEM (accessed April 15, 2022).

26. Wolfinge R, Lander D. *LockNQuay - LNQ-ens1*. (2020). Available online at: https://www.kaggle.com/sasrdw/locknquay (accessed April 15, 2022).

27. Ray EL, Tibshirani R. *COVIDhub-Baseline*. (2020). Available online at: https://github.com/reichlab/covidModels (accessed April 15, 2022).

28. Ray EL, Cramer E, Gerding A, Reich N. *COVIDhub-trained_ensemble*. (2021). Available online at: https://github.com/reichlab/covid19-forecast-hub (accessed April 15, 2022).

29. Adiga A, Wang L, Venkatramanan S, Peddireddy AS, Hurt B, Lewis B, et al. *UVA-Ensemble*. (2020). Available online at: https://biocomplexity.virginia.edu/ (accessed April 15, 2022).

30. Perakis G, Spantidakis I, Tsiourvas A, Ndong DN, Bennouna M, Thayaparan L, et al. *MIT-Cassandra*. (2021). Available online at: https://github.com/oskali/mit_cassandra (accessed April 15, 2022).

31. Yogurtcu ON, Messan MR, Gerkin RC, Belov AA, Yang H, Forshee RA, et al. A quantitative evaluation of COVID-19 epidemiological models. *medRxiv* [preprint]. (2021). doi: 10.1101/2021.02.06.21251276

32. Kinsey M, Tallaksen K, Obrecht RF, Asher L, Costello C, Kelbaugh M, et al. *JHUAPL-Bucky*. (2020). Available online at: https://github.com/mattkinsey/bucky (accessed April 15, 2022).

33. Wilson DJ. Weather, social distancing, and the spread of COVID-19. In: *Federal Reserve Bank of San Francisco, Working Paper Series*. (2020). p. 1–64. Available online at: https://www.medrxiv.org/content/10.1101/2020.07.23.20160911v1 (accessed April 15, 2022).

34. Suchoski B, Stage S, Gurung H, Baccam S. *IEM_MED-CovidProject*. (2020). Available online at: https://iem-modeling.com/ (accessed April 15, 2022).

35. Vespignani A, Chinazzi M, Davis JT, Mu K, Pastore y Piontti A, Samay N, et al. *MOBS-GLEAM_COVID*. (2020). Available online at: https://uploads-ssl.webflow.com/58e6558acc00ee8e4536c1f5/5e8bab44f5baae4c1c2a75d2_GLEAM_web.pdf (accessed April 15, 2022).

36. Kalantari R, Zhou M. *DDS-NBDS*. (2020). Available online at: https://dds-covid19.github.io/ (accessed April 15, 2022).

37. Jain CL. *IBF-TimeSeries*. (2020). Available online at: https://ibf.org/ (accessed April 15, 2022).

38. Walraven R. *RobertWalraven-ESG*. (2020). Available online at: http://rwalraven.com/COVID19 (accessed April 15, 2022).

39. Corsetti S, Myers R, Miao J, Huang Y, Falb K, Schwarz T. *UMich-RidgeTfReg*. (2020). Available online at: https://gitlab.com/sabcorse/covid-19-collaboration (accessed April 15, 2022).

40. Karlen D. Karlen-pypm: Characterizing the spread of CoViD-19. *arXiv* [preprint]. (2020). Available online at: doi: 10.48550/arXiv.2007.07156

41. Osthus D, Del Valle S, Manore C, Weaver B, Castro L, Shelley C, et al. *LANL-GrowthRate*. (2020). Available online at: https://covid-19.bsvgateway.org/ (accessed April 15, 2022).

42. Burant J. *JCB-PRM*. (2020). Available online at: https://github.com/JohnBurant/COVID19-PRM (accessed April 15, 2022).

43. Nagraj VP, Hulme-Lowe C, Guertin SL, Turner SD. FOCUS: Forecasting COVID-19 in the United States. *medRxiv*. [preprint]. (2021). doi: 10.1101/2021.05.18.21257386

44. O'Dea E. *CEID-Walk* (2020). Available online at: https://github.com/e3bo/random-walks (accessed April 15, 2022).

45. Jadbabaie A, Sarker A, Shah D. *MIT_ISOLAT-Mixtures*. (2020). Available online at: https://idss.mit.edu/vignette/real-time-mixture-based-predictions/ (accessed April 15, 2022).

46. Schwanke Khilji SU, Rudge JW, Drake T, Chavez I, Borin K, Touch S, et al. Distribution of selected healthcare resources for influenza pandemic response in Cambodia. *Int J Equity Health*. (2013) 12:82. doi: 10.1186/1475-9276-12-82

47. Rudge JW, Hanvoravongchai P, Krumkamp R, Chavez I, Adisasmito W, Ngoc Chau P, et al. Health system resource gaps and associated mortality from pandemic influenza across six Asian territories. *PLoS ONE*. (2012) 7:e31800. doi: 10.1371/journal.pone.0031800

48. Worby CJ, Chang HH. Face mask use in the general population and optimal resource allocation during the COVID-19 pandemic. *Nat Commun*. (2020) 11:404. doi: 10.1038/s41467-020-17922-x

49. Nandakishore P, Liu M, R P, Gourneni S, Sukumaran R, Berman JM, et al. Deviations in predicted COVID-19 cases in the US during early months of 2021 relate to rise in B.1.526 and its family of variants. *medRxiv* [preprint]. (2021). doi: 10.1101/2021.12.06.21267388

50. Drake RG. Does longer roster lead-time reduce temporary staff usage? A regression analysis of e-rostering data from 77 hospital units. *J Adv Nurs*. (2018) 74:1831–8. doi: 10.1111/jan.13578

51. Bonnet L, Carle A, Muret J. In the light of COVID-19 oxygen crisis, why should we optimise our oxygen use? *Anaesth Crit Care Pain Med*. (2021) 40:100932. doi: 10.1016/j.accpm.2021.100932

52. Bhuyan A. Experts criticise India's complacency over COVID-19. *Lancet*. (2021) 397:1611–2. doi: 10.1016/S0140-6736(21)00993-4