Harnessing the Wisdom of the Crowd to Forecast Incident and Cumulative COVID-19 Mortality in the United States

Kathryn S. Taylor¹,* James W Taylor²
* Corresponding author kathryn.taylor@phc.ox.ac.uk

Affiliations
1Nuffield Department of Primary Care Health Sciences, University of Oxford
2Saïd Business School, University of Oxford

Abstract

Background
Forecasting models have played a pivotal role in decision making during the COVID-19 pandemic, predicting the numbers of cases, hospitalisations and deaths. However, questions have been raised about the role and reliability of models. The aim of this study was to investigate the potential benefits of combining probabilistic forecasts from multiple models for forecasts of incident and cumulative COVID mortalities.

Methods
We considered 95% interval and point forecasts of weekly incident and cumulative COVID-19 mortality between 16 May 2020 and 8 May 2021 in multiple locations in the United States. We compared the accuracy of simple and more complex combining methods, as well as individual models.

Results
The average of the forecasts from the individual models was consistently more accurate than the average performance of these models, which provides a fundamental motivation for combining. Weighted combining performed well for both incident and cumulative mortalities, and for both interval and point forecasting. Inverse score with tuning was the most accurate method overall. The median combination was a leading method in the last quarter for both mortalities, and it was consistently more accurate than the mean combination for point forecasting of both mortalities. For interval forecasts of cumulative mortality, the mean performed better than the median. The leading individual models were most competitive for point forecasts of incident mortality.

Conclusions
We recommend that harnessing the wisdom of the crowd can improve the contribution of probabilistic forecasting of epidemics to health policy decision making, and report that the relative performance of the different combining methods depends on several factors including the type of data, type of forecast and timing.

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
Introduction

The coronavirus disease-2019 (COVID-19) pandemic has created major planning and resource allocation challenges, as well as pressures on health services, which have prompted governments to impose extreme restrictions in attempts to control the spread of the virus (1-3). These measures have resulted in multiple problems beyond COVID-19, including increased hospital treatment delays, damage to economies, higher levels of unemployment, declining mental health, and widening of the pre-existing health and educational disparities, which will persist beyond the rollout of vaccines (4, 5). Subsequently, the pandemic has generated intense debate among experts about the best way forward (6). Governments and their advisors have relied upon forecasts from models of the numbers of COVID-19 cases, people hospitalized and deaths to help decide what actions to take (7). However, using modelling to lead health policy during the pandemic has been controversial and criticised on various grounds, including the overreliance on models and questionable model assumptions. Nevertheless, it is recognised that models are potentially valuable tools when used appropriately (2, 8-10).

The most useful models are parsimonious, that is, including only sufficient detail to answer a particular question (11), and parameterised with evidence-based data, rather than being based on assumptions. Using models that provide frequent forecasts will incorporate the latest evidence into the model estimates, as well as realign with the latest mitigating measures by governments and the responses of their populations. Furthermore, forecasts will be more nuanced if modelling is carried out at the local, rather than national level (2). Models may be constructed for prediction or scenario analysis. For example, the model from Imperial College London that drove the United Kingdom and United States (U.S.) governments to impose the first lockdowns assumed scenarios where between 50% and 75% of people would comply with the government restrictions (12). Extreme assumptions may be made to provide policy insight into a broad set of possible scenarios (13). When employed for prediction, models forecast the most likely outcome in the current circumstances. Different models are based on different approaches to answering the same question, and conflicting forecasts may arise. Rather than questioning which model is best (13), a forecast combination can be used to harness the wisdom of the crowd (14). Combining produces a collective forecast from multiple models that is typically more accurate than forecasts from individual models. The mean combination (simple average of all the forecasts) is an example of a combined forecast, which is often used and hard to beat (15, 16). Forecast combining has many advantages (17). It synthesises information underlying different prediction methods in a pragmatic way, diversifying the risk inherent in relying on an individual model, and it can offset the statistical bias associated with individual models, potentially cancelling out overestimation and underestimation from individual models. In many applications, from economics and business (18, 19), to weather and climate prediction (20, 21), the advantages of forecast combining are well-established. This has encouraged the more recent application of combining to the prediction of infectious diseases (22-26). When considering forecasting, attention is often placed on point forecasts, but they have inherent uncertainty, and subsequently, there has been increasing calls for probabilistic forecasting (9, 22). An interval forecast is a common form of probabilistic forecast, which conveys the uncertainty in an intuitively appealing way (27). For example, a 95% interval forecast is a range that will, ideally, contain the true value with 95% probability.
Previously, we compared the accuracy of probabilistic forecast combining methods applied to predictions of weekly cumulative COVID-19 mortality in U.S. locations over the 40-week period up to 23 January 2021 (26). In this paper, we extend our earlier study by considering both weekly cumulative and incident COVID-19 mortalities, where incident deaths for a week is defined as the number of deaths occurring in that week. The numbers of cumulative deaths and incident deaths are clearly related, but although observed values of one can be derived from the other, the same is not true for probabilistic forecasts beyond one step-ahead. Furthermore, different sets of models may be used to predict incident and cumulative mortality. We consider forecast combining for point forecasts and 95% interval forecasts for the 52-week period up to 15 May 2021, which constitutes a longer and more recent period of data than in our previous work. Additional differences from our earlier study are that we report the results for individual models in our comparison, and assess the impact of reporting delays of death counts on forecast accuracy.

COVID-19 mortality dataset

Data sources

Forecasts of weekly incident and cumulative COVID-19 mortalities were downloaded from the COVID-19 Forecast Hub (23). The Hub is an ongoing collaboration between the U.S. Centers for Disease Control and Prevention (CDC), with forecasting teams from academia, industry and government-affiliated groups (28). Teams are invited to submit forecasts of incident and cumulative mortalities for 1, 2, 3 and 4-week horizons, in the form of estimates of quantiles corresponding to 23 probability points along the probability distribution of possible values, and also point forecasts. Forecasts of the 2.5% and 97.5% quantiles bound a 95% interval forecast. In this paper, in addition to 95% interval forecasts, we consider point forecasts, which, for each model, we obtain as the 50% quantile (i.e., the median) at the centre of the probability distribution. The numbers of actual incident COVID-19 deaths each week were inferred from the reference data from the Hub on the numbers of actual weekly cumulative COVID-19 deaths. These data are provided by the Centre for Systems Science and Engineering (CSSE) at John Hopkins University (28).

Analysis dataset

The dataset for our analysis included forecasts projected from forecast origins each week between the weeks ending Saturday 16 May 2020 and Saturday 8 May 2021 (52 weeks of data). The COVID-19 Forecast Hub numbers these weeks as Epidemic Weeks 21 to 72, where Week 0 was the week ending 21 December 2019. These forecasts were compared with actual weekly COVID-19 mortality up to the week ending 15 May 2021 (Week 73). We studied forecasts of COVID-19 mortality for the whole of the U.S. and 51 U.S. jurisdictions, which included the 50 states and the District of Columbia. For simplicity, in the rest of the paper, we refer to the jurisdictions as states. For each state, Figs. S1 and S2 in the supplementary information show the numbers of weekly incident and cumulative COVID-19 deaths respectively, from Weeks 22 to 73. These figures show different histories of the pandemic across the states.
The Hub carries out various screening tests for inclusion in their forecast combination, which they refer to as their ‘ensemble’ model. We included forecasts that passed the Hub’s screening tests, as well as forecasts that were not submitted in time to be screened. We followed the Hub by excluding, for any given week, a model for which forecasts for all 23 quantiles and for all four forecast horizons were not provided. The Hub also excluded forecasts deemed to be outlying. We did not exclude outliers, primarily because the actual number of COVID-19 deaths in previous weeks may have been updated, and therefore the assessment of outliers in the past, by the Hub, would not be consistent with our retrospective assessments of outliers at Week 73.

Delays in reporting COVID-19 deaths is a well-recognised problem. Updates, typically increases, which may involve sharp increases in death counts, will result in forecasting models underestimating, and when updates are backdated, this will lead to forecasting methods being penalised in retrospective analyses of forecast accuracy. Updates that decrease death counts will produce overestimates. We downloaded 13 data files of actual death counts at different points during our 53-week study period (one file each at Weeks 26, 30, 32, 34, 37, 41, 43, 45, 47, 54, 59, 66 and 73). We observed updates to the historical death counts in 14 locations between Weeks 26 and 73 (Alaska, California, Delaware, Indiana, Ohio, Oklahoma, Missouri, New Jersey, New York, Rhode Island, Texas, Washington, Wisconsin and the U.S.). Fig. 1 presents superimposed plots of the numbers of cumulative COVID-19 deaths, using the 13 data files, for six states in which the effects of updates were particularly notable. We evaluate the effect of reporting delays on forecast accuracy when we compare the forecasting methods.

![Fig. 1 Numbers of reported deaths in states where there were notable effects of reporting delays: (a) Delaware, (b) Indiana, (c) Ohio, (d) Rhode Island, (e) Oklahoma and (f) West Virginia. Updates with backdating (a-d) are shown as deviations (dashed lines) from the most recent plot of cumulative COVID-19 mortality reported at Week 73 (black lines), and updates without backdating are shown as a sharp upward step (e) and a decrease (f) between Weeks 66 and 73.](mediRxiv preprint: https://doi.org/10.1101/2021.07.11.21260318; this version posted July 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.)
Our analysis included forecasts from 53 individual forecasting models and the Hub’s ensemble model. Details about these models are given in the appendix in the supplementary information. In the early weeks of our dataset, the traditional SEIR (susceptible-exposed-infected-removed) compartmental models were in the majority, but as the weeks passed, other types of models became more common (Fig. 2). These other models involved various statistical techniques and methods including neural networks, agent-based models, time series modelling, ridge regression and curve fitting techniques.

![Fig. 2 Number and types of models at each forecast origin for each mortality dataset](image)

Fig. 2 Number and types of models at each forecast origin for each mortality dataset

Fig. 3 shows the timeline of forecasts from each model, illustrating the extent of missing data across the 52 locations, including the frequent ‘entry and exit’ of forecasting teams. This figure also shows that there were differences between the sets of models providing forecasts of incident and cumulative mortality. Screening by the Hub removed 14.4% of forecasts of incident deaths and 22.3% of forecasts of cumulative deaths. The large amount of missing data was such that we felt imputation would not be appropriate.

![Fig. 3 Extent of missing data for forecasts of incident mortality (left) and cumulative mortality (right) from Models 1 to 53 for forecast origins between Epidemic Weeks 21 and 72](image)

Fig. 3 Extent of missing data for forecasts of incident mortality (left) and cumulative mortality (right) from Models 1 to 53 for forecast origins between Epidemic Weeks 21 and 72
Comparison of forecast accuracy

Forecast evaluation methods

The accuracy of the 95% interval forecasts was evaluated in terms of calibration and the interval score. Calibration was assessed by the percentage of actual deaths that fell within the bounds of the interval forecasts, with the ideal being 95%. The interval score was calculated by the following expression (29, 30):

\[ S^\text{INT}_t(l_t, u_t, y_t) = (u_t - l_t) + \frac{2}{\alpha} I\{y_t \leq l_t\} (l_t - y_t) + \frac{2}{\alpha} I\{y_t \geq u_t\} (y_t - u_t) \]

where \( l_t \) is the interval’s lower bound, \( u_t \) is its upper bound, \( y_t \) is the observation in period \( t \), \( I \) is the indicator function (1 if the condition is true and 0 otherwise), and, for a 95% interval, \( \alpha=5\% \). The bounds \( l_t \) and \( u_t \) are the 2.5% and 97.5% quantiles, respectively. Lower values of the interval score reflect greater interval forecast accuracy, and for this study, the unit of the quantile score is the number of deaths. The score rewards narrow intervals, with observations that fall outside the interval incurring a penalty, the magnitude of which depends on the value of \( \alpha \) (29). The accuracy of the point forecasts was evaluated using the absolute value of the forecast errors. Averaging each of these two scores across an out-of-sample period provides two measures of forecast accuracy – the mean absolute error (MAE) and the mean interval score (MIS).

In most of our reporting of the results, we average across horizons. We do this for conciseness and because we are using a relatively short out-of-sample period, which is a particular problem when evaluating forecasts of extreme quantiles. To show the consistency across horizons, we present results by horizon for interval forecasts. For these MIS results, we carried out statistical tests for the difference in forecast accuracy of different methods and models. We adapted the Diebold-Mariano test (31) in order to test across multiple series, and this was applied to the results of each prediction horizon separately. To summarise results averaged across the four horizons, we were unable to use the Diebold-Mariano test, so we applied the statistical test proposed by Koning et al. (32), which, for each method, compares the rank, averaged across multiple series, with the corresponding average rank of the most accurate method. Statistical testing was based on a 5% significance level.

Interval forecast combining methods

The comparison included several combining methods that do not rely on the availability of records of past accuracy. These methods are useful in the early stages of a pandemic, and in later stages when a new forecasting team starts to submit forecasts, or there is an uneven record of past accuracy among the different models, which is the case in our study. Fig. 4 presents a visual summary of these combining methods. Combining is applied to each interval bound separately. The mean combination (a, in Fig.4) and the median combinations (b, in Fig.4) are two well established benchmark methods (33-35). The median has the appeal of being robust to outliers.
Fig. 4 Illustration of non-score-based combining methods of interval forecasts from five individual models, showing (a) mean combination, (b) median combination, (c) symmetric trimmed mean, (d) asymmetric exterior trimmed mean, (e) asymmetric interior trimmed mean and (f) envelope. Each interval forecast is represented by a different coloured pair of dots.

More novel methods of combining involve trimming (excluding) a particular percentage $\beta$ of forecasts of each interval bound (shown by shading in Fig. 4), and then averaging the remaining forecasts of that bound. **Symmetric trimming** (c, in Fig.4) deals with outliers. For each bound, it involves trimming the $N$ lowest-valued and $N$ highest-valued forecasts, where $N$ is the largest integer less than or equal to the product of $\beta/2$ and the total number of forecasts (36). The median combination is an extreme form of symmetric trimming. We also implemented asymmetric exterior trimming, asymmetric interior trimming, and the envelope methods (33). **Asymmetric exterior trimming** (d, in Fig.4) is suitable for addressing underconfidence, which is reflected by overly wide intervals. It involves removing the $N$ lowest-valued lower bound forecasts, as well as the $N$ highest-valued upper bound forecasts, where $N$ is the largest integer less than or equal to the product of $\beta$ and the number of forecasts. When trimming resulted in a lower bound being above the upper bound, we replaced the two bounds by their average. **Asymmetric interior trimming** (e, in Fig.4) deals with overconfidence, which is reflected by overly narrow intervals. For this, we removed the $N$ highest-valued lower bound forecasts and the $N$ lowest-valued upper bound forecasts, where $N$ is defined as for exterior trimming. The **envelope** method (f, in Fig.4) is an extreme form of interior trimming, whereby the interval is constructed using the lowest-valued lower bound forecast and highest-valued upper bound forecast.

For our comparison, we divided the 52 weeks, from which forecasts were made, into an expanding in-sample period, starting with length 13 weeks, and a 39-week out-of-sample period. For each week, location and method, we optimised the trimming percentage $\beta$ by finding the value that minimised the sum of the MIS averaged over all four horizons and all periods up to and including the latest forecast origin. Optimisation occurred each week, using the expanding in-sample period, to produce combined forecasts for the 39 weeks of the out-of-sample period.

We included the COVID-19 Hub’s *ensemble* forecast. This was initially the mean combination of the forecasts that they considered to be eligible, but in late July, 12 weeks into our 52-week analysis period, it
became the median combination of these forecasts (23). The use of eligibility screening implies that the ensemble is constructed with the benefit of a degree of subjective trimming. The ensemble provided forecasts of cumulative mortality for all of our 52-week period, while for incident mortality, forecasts were available from all except the first three weeks of the 52-week period.

In our comparison of methods, we also included three methods that considered the accuracy of the forecasts from the individual models. The first of these three methods simply selected the model with the best historical accuracy (37). We refer to this method as previous best. For this method, the interval forecast was obtained from the model for which the MIS was the lowest when computed using the weeks up to and including the current forecast origin (i.e., the in-sample period). We also investigated two weighted average combinations, which have similarities with inverse-variance weighting, which is a common approach used in meta-analysis (38). For one weighted method, the inverse interval score method, the weights were inversely proportional to the MIS (26). For the other weighted method, inverse interval score with tuning, a tuning parameter, $\lambda > 0$, is incorporated to control the influence of the score on the combining weights, using the following expression for the weight on forecasting model $i$ at forecast origin $t$ (26):

$$w_{it} = \frac{(1/MIS_{it})^{\lambda}}{\sum_{j=1}^{J}(1/MIS_{jt})^{\lambda}}$$

where $MIS_{it}$ is the historical MIS computed at forecast origin $t$ from model $i$, and $J$ is the number of forecasting models included in the combination. If $\lambda$ is close to zero, the combination reduces to the mean combination, whereas a large value for $\lambda$ leads to the selection of the model with best historical accuracy. The parameter $\lambda$ was optimised using the same expanding in-sample periods, as for the trimming combining methods. Due to the extent of missing forecasts (Fig.3), we pragmatically computed $MIS_{it}$ using all available past forecasts, rather than limit the computation to periods for which forecasts from all models were available. For the models for which forecasts were not available for at least 5 past periods, we set $MIS_{it}$ to be equal to the mean of $MIS_{it}$ for all other models. An alternative approach, employed in (26), is to omit from the combination any model for which there is only a very short or non-existent history of accuracy available. The disadvantage of this is that it omits potentially useful forecast information, and this was supported by our empirical forecasting results.

**Point forecast combining methods**

For the point forecasts, we considered analogous combining methods to those for the interval forecasts, with the exception of the asymmetric interior trimming, asymmetric exterior trimming and envelope combining methods, which are only of use for interval forecast combining.

**Inclusion of individual models in the comparison**

Alongside the results for the combining methods, we also report the results for any individual model for which forecasts were available for all 52 locations and all 39 out-of-sample periods. There were six models fulfilling this criterion for incident mortality (numbered 1, 14, 20, 21, 22 and 33), and two models for cumulative mortality (numbered 21 and 33). The numbering of models corresponds to that shown in Fig.3.
Empirical results

Performance of methods overall

We averaged the MIS and MAE across all four horizons and all 52 locations: the 51 states and the whole U.S. As the unit of the two scores was deaths, to avoid scores for some locations dominating, we also present results averaged for the following four categories: high, medium and low mortality states, and the U.S. as a whole. The high, medium and low categories were decided by ordering the 51 states by the number of cumulative deaths at the end of the final week of our dataset, and then dividing the states into three groups of 17 states. Tables 1 and 2 present the MIS and mean ranks for 95% interval forecasts, for each method for the 32-week out-of-sample period for incident and cumulative mortality, respectively. Tables 3 and 4 present the corresponding MAE results for point forecasts.

Considering methods that performed well in terms of either being the leading method or competitive against the leading method, inverse score combining performed best overall, with the method benefitting from the incorporation of tuning (Tables 1 to 4). The poorest results were produced by the envelope method (Tables 1 to 4). The asymmetric interior trimmed mean performed well for interval forecasting for low and medium mortality states, producing the most accurate interval forecasts for the low mortality category for both the incident and cumulative mortality data (Tables 1 and 2). The leading individual model (Model 33) performed well in forecasting the numbers of deaths for all series, the U.S. as a whole, and for high mortality states (Tables 1 to 4). Its performance was stronger in point forecasting than interval forecasting, and it was more accurate in forecasting incident deaths than cumulative deaths (Tables 1 to 4).

The final row in each of Tables 1 to 4 does not provide the forecasting results for a method. Instead, it reports the average score of all the individual methods. In each table, we see that this is substantially worse than the performance of the mean combining method (i.e., the average performance is much worse than the performance of the average). This gives fundamental support for forecast combining, because the average performance can be viewed as the statistical expectation for the performance of an individual model, chosen when there is no information regarding their accuracy, which was the case at the start of the pandemic.

Comparing the two simple combining methods, for interval forecasts of incident deaths, the median was more accurate than the mean for all series and for high mortality states, whilst the performance was similar for medium and low mortality states (Table 1). For interval forecasts of cumulative mortality, the mean was more accurate than the median across all categories (Table 2). The median was consistently more accurate than the mean for point forecasts of both cumulative and incident mortality (Tables 3 and 4).
Table 1 For 95% interval forecasts of incident mortality, MIS and mean ranks for the 39-week out-of-sample period. Boxed numbers indicate the best method in each column.

<table>
<thead>
<tr>
<th>Method</th>
<th>MIS</th>
<th>Mean rank</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All</td>
<td>U.S.</td>
</tr>
<tr>
<td>Mean</td>
<td>900</td>
<td>9799</td>
</tr>
<tr>
<td>Median</td>
<td>778</td>
<td>11239</td>
</tr>
<tr>
<td>Ensemble</td>
<td>794</td>
<td>11176</td>
</tr>
<tr>
<td>Sym trim</td>
<td>878</td>
<td>10445</td>
</tr>
<tr>
<td>Asym exterior trim</td>
<td>996</td>
<td>12447</td>
</tr>
<tr>
<td>Asym interior trim</td>
<td>895</td>
<td>9969</td>
</tr>
<tr>
<td>Envelope</td>
<td>5284</td>
<td>64944</td>
</tr>
<tr>
<td>Previous best</td>
<td>877</td>
<td>11779</td>
</tr>
<tr>
<td>Inverse score</td>
<td>770</td>
<td>9130</td>
</tr>
<tr>
<td>Inverse score tuning</td>
<td>677</td>
<td>8939</td>
</tr>
<tr>
<td>Model 1</td>
<td>729</td>
<td>11174</td>
</tr>
<tr>
<td>Model 14</td>
<td>1918</td>
<td>30502</td>
</tr>
<tr>
<td>Model 20</td>
<td>1957</td>
<td>33948</td>
</tr>
<tr>
<td>Model 21</td>
<td>2004</td>
<td>35399</td>
</tr>
<tr>
<td>Model 22</td>
<td>1014</td>
<td>19145</td>
</tr>
<tr>
<td>Model 33</td>
<td>748</td>
<td>8681</td>
</tr>
<tr>
<td>Mean MIS of models</td>
<td>2222</td>
<td>37870</td>
</tr>
</tbody>
</table>

* The method is significantly worse than the method with the best mean rank. MIS – Mean interval score.

Table 2 For 95% interval forecasts of cumulative mortality, MIS and mean ranks for the 39-week out-of-sample period. Boxed numbers indicate the best method in each column.

<table>
<thead>
<tr>
<th>Method</th>
<th>MIS</th>
<th>Mean rank</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All</td>
<td>U.S.</td>
</tr>
<tr>
<td>Mean</td>
<td>3529</td>
<td>48497</td>
</tr>
<tr>
<td>Median</td>
<td>3700</td>
<td>56185</td>
</tr>
<tr>
<td>Ensemble</td>
<td>3588</td>
<td>50481</td>
</tr>
<tr>
<td>Sym trim</td>
<td>3671</td>
<td>52563</td>
</tr>
<tr>
<td>Asym exterior trim</td>
<td>3737</td>
<td>53839</td>
</tr>
<tr>
<td>Asym interior trim</td>
<td>3883</td>
<td>61412</td>
</tr>
<tr>
<td>Envelope</td>
<td>6545</td>
<td>103456</td>
</tr>
<tr>
<td>Previous best</td>
<td>3531</td>
<td>49146</td>
</tr>
<tr>
<td>Inverse score</td>
<td>3296</td>
<td>40479</td>
</tr>
<tr>
<td>Inverse score tuning</td>
<td>3100</td>
<td>40758</td>
</tr>
<tr>
<td>Model 21</td>
<td>7247</td>
<td>136492</td>
</tr>
<tr>
<td>Model 33</td>
<td>3457</td>
<td>44297</td>
</tr>
<tr>
<td>Mean MIS of models</td>
<td>6486</td>
<td>114911</td>
</tr>
</tbody>
</table>

* The method is significantly worse than the method with the best mean rank. MIS – Mean interval score.
Table 3  For point forecasts of incident mortality, MAE and mean ranks for the 39-week out-of-sample period. Boxed numbers indicate the best method in each column.

<table>
<thead>
<tr>
<th>Method</th>
<th>All</th>
<th>U.S.</th>
<th>High</th>
<th>Med</th>
<th>Low</th>
<th>All</th>
<th>U.S.</th>
<th>High</th>
<th>Med</th>
<th>Low</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>99</td>
<td>1711</td>
<td>146</td>
<td>43</td>
<td>13</td>
<td>6.1</td>
<td>8</td>
<td>6.5</td>
<td>5.5</td>
<td>6.1</td>
</tr>
<tr>
<td>Median</td>
<td>86</td>
<td>1632</td>
<td>114</td>
<td>41</td>
<td>13</td>
<td>4.2</td>
<td>6</td>
<td>4.6</td>
<td>3.8</td>
<td>4.1</td>
</tr>
<tr>
<td>Ensemble</td>
<td>87</td>
<td>1630</td>
<td>114</td>
<td>41</td>
<td>13</td>
<td>4.8</td>
<td>5</td>
<td>4.8</td>
<td>4.2</td>
<td>5.3</td>
</tr>
<tr>
<td>Sym trim</td>
<td>95</td>
<td>1632</td>
<td>141</td>
<td>41</td>
<td>13</td>
<td>4.9</td>
<td>7</td>
<td>5.4</td>
<td>4.5</td>
<td>4.6</td>
</tr>
<tr>
<td>Previous best</td>
<td>100</td>
<td>1889</td>
<td>130</td>
<td>48</td>
<td>16</td>
<td>8.4*</td>
<td>9</td>
<td>7.7</td>
<td>8.5*</td>
<td>8.9*</td>
</tr>
<tr>
<td>Inverse score</td>
<td>95</td>
<td>1585</td>
<td>143</td>
<td>42</td>
<td>13</td>
<td>4.9</td>
<td>4</td>
<td>5.5</td>
<td>4.9</td>
<td>4.5</td>
</tr>
<tr>
<td>Inverse score tuning</td>
<td>82</td>
<td>1356</td>
<td>116</td>
<td>43</td>
<td>13</td>
<td>5.4</td>
<td>2</td>
<td>4.8</td>
<td>5.6</td>
<td>6.1</td>
</tr>
<tr>
<td>Model 1</td>
<td>83</td>
<td>1460</td>
<td>112</td>
<td>44</td>
<td>13</td>
<td>10.4*</td>
<td>13</td>
<td>10.8*</td>
<td>10.3*</td>
<td>9.9*</td>
</tr>
<tr>
<td>Model 14</td>
<td>114</td>
<td>2097</td>
<td>155</td>
<td>55</td>
<td>16</td>
<td>4.6</td>
<td>3</td>
<td>5.4</td>
<td>6.1</td>
<td>2.5</td>
</tr>
<tr>
<td>Model 20</td>
<td>117</td>
<td>2216</td>
<td>155</td>
<td>56</td>
<td>17</td>
<td>10.7*</td>
<td>10</td>
<td>10.3*</td>
<td>11.1*</td>
<td>10.8*</td>
</tr>
<tr>
<td>Model 21</td>
<td>118</td>
<td>2247</td>
<td>157</td>
<td>56</td>
<td>17</td>
<td>10.5*</td>
<td>11</td>
<td>10.5*</td>
<td>10.4*</td>
<td>10.7*</td>
</tr>
<tr>
<td>Model 22</td>
<td>119</td>
<td>2401</td>
<td>151</td>
<td>55</td>
<td>16</td>
<td>11.1*</td>
<td>12</td>
<td>11.1*</td>
<td>11.1*</td>
<td>11.2*</td>
</tr>
<tr>
<td>Model 33</td>
<td>80</td>
<td>1277</td>
<td>114</td>
<td>42</td>
<td>13</td>
<td>4.8</td>
<td>1</td>
<td>3.5</td>
<td>5.0</td>
<td>6.3</td>
</tr>
<tr>
<td>Mean MAE of models</td>
<td>130</td>
<td>2309</td>
<td>188</td>
<td>57</td>
<td>17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* The method is significantly worse than the method with the best mean rank. MAE – Mean absolute error.

Table 4 For point forecasts of cumulative mortality, MAE and mean ranks for the 39-week out-of-sample period. Boxed numbers indicate the best method in each column.

<table>
<thead>
<tr>
<th>Method</th>
<th>All</th>
<th>U.S.</th>
<th>High</th>
<th>Med</th>
<th>Low</th>
<th>All</th>
<th>U.S.</th>
<th>High</th>
<th>Med</th>
<th>Low</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>270</td>
<td>4754</td>
<td>430</td>
<td>87</td>
<td>30</td>
<td>5.5</td>
<td>8</td>
<td>5.6</td>
<td>5.2</td>
<td>5.4</td>
</tr>
<tr>
<td>Median</td>
<td>240</td>
<td>4438</td>
<td>359</td>
<td>84</td>
<td>30</td>
<td>4.4</td>
<td>5</td>
<td>5.1</td>
<td>3.8</td>
<td>4.3</td>
</tr>
<tr>
<td>Ensemble</td>
<td>238</td>
<td>4373</td>
<td>357</td>
<td>85</td>
<td>30</td>
<td>4.8</td>
<td>4</td>
<td>4.5</td>
<td>4.3</td>
<td>5.8</td>
</tr>
<tr>
<td>Sym trim</td>
<td>244</td>
<td>4639</td>
<td>359</td>
<td>84</td>
<td>29</td>
<td>4.5</td>
<td>7</td>
<td>5.5</td>
<td>4.3</td>
<td>3.5</td>
</tr>
<tr>
<td>Previous best</td>
<td>253</td>
<td>3976</td>
<td>398</td>
<td>108</td>
<td>32</td>
<td>7.5*</td>
<td>2</td>
<td>6.7*</td>
<td>8.6*</td>
<td>7.4*</td>
</tr>
<tr>
<td>Inverse score</td>
<td>262</td>
<td>4576</td>
<td>417</td>
<td>87</td>
<td>29</td>
<td>4.8</td>
<td>6</td>
<td>5.2</td>
<td>5.0</td>
<td>4.1</td>
</tr>
<tr>
<td>Inverse score tuning</td>
<td>236</td>
<td>4291</td>
<td>353</td>
<td>86</td>
<td>29</td>
<td>4.3</td>
<td>3</td>
<td>4.4</td>
<td>4.6</td>
<td>4.0</td>
</tr>
<tr>
<td>Model 21</td>
<td>346</td>
<td>6914</td>
<td>480</td>
<td>128</td>
<td>42</td>
<td>9.6*</td>
<td>10</td>
<td>9.6*</td>
<td>9.5*</td>
<td>9.7*</td>
</tr>
<tr>
<td>Model 33</td>
<td>231</td>
<td>3910</td>
<td>360</td>
<td>88</td>
<td>30</td>
<td>4.7</td>
<td>1</td>
<td>3.0</td>
<td>4.8</td>
<td>6.4</td>
</tr>
<tr>
<td>Mean MAE of models</td>
<td>338</td>
<td>6102</td>
<td>515</td>
<td>120</td>
<td>38</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* The method is significantly worse than the method with the best mean rank. MAE – Mean absolute error.
Tables S1 and S2 in the supplementary information provide a broad summary of the results of Tables 1 to 4 in terms of the frequency with which a method is ranked in the top three, based on the scores and mean ranks, respectively, for the different types of mortality and forecast. These tables reflect the dominance of the inverse score methods, for both mortalities and both forecast types, although the superiority was less for point forecasts. The inverse score with tuning is best overall according to both the scores and mean ranks.

Looking at the results for the statistical testing of the mean ranks, we see that the methods that performed poorly were identified as being statistically significantly worse than the best methods. Tables 1 to 4 report results averaged across the four forecast horizons (1 to 4 weeks ahead). We found similar relative performances of the methods when looking at each forecast horizon separately (see Tables S3 and S4 in the supplementary information).

**Changes in performance over time**

Tables 5 and 6 show the MIS and MAE for the four quarters of our 1-year sample of data. Note that, for the first 13-week period, results were not obtainable for the methods for which a previous in-sample period was needed to estimate method parameters. The relative performances of the methods in Tables 5 and 6 are reasonably consistent with the results in Tables 1 to 4, where we averaged across the 39-week out-of-sample period. In Tables 5 and 6, we note the much higher values of the scores in the third quarter, caused by the much higher levels of mortality during that period. It is interesting to see from Tables 5 and 6 that, for forecasts of both incident and cumulative mortalities, the leading methods in the most recent 13-week period were the median combination and the ensemble, which, after the first quarter, was computed using the median. In the second and third quarters, the mean was more accurate than the median for both incident and cumulative mortalities.
Table 5 For incident mortality, scores for each quarter of our 52-week dataset. The unit of the scores is deaths. For each quarter, boxed numbers indicate the best method for each column.

<table>
<thead>
<tr>
<th>Method</th>
<th>All</th>
<th>U.S.</th>
<th>MIS</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>High</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Med</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Low</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First 13-week period</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>503</td>
<td>5992</td>
<td>978</td>
<td>166</td>
</tr>
<tr>
<td>Median</td>
<td>419</td>
<td>6297</td>
<td>728</td>
<td>150</td>
</tr>
<tr>
<td>Model 1</td>
<td>450</td>
<td>7226</td>
<td>703</td>
<td>199</td>
</tr>
<tr>
<td>Previous best</td>
<td>56</td>
<td>1064</td>
<td>81</td>
<td>23</td>
</tr>
<tr>
<td>Ensemble</td>
<td>48</td>
<td>821</td>
<td>73</td>
<td>22</td>
</tr>
<tr>
<td>Mean</td>
<td>33</td>
<td>1</td>
<td>51</td>
<td>419</td>
</tr>
<tr>
<td>Median</td>
<td>50</td>
<td>172</td>
<td>48</td>
<td>452</td>
</tr>
<tr>
<td>Inverse score</td>
<td>50</td>
<td>172</td>
<td>48</td>
<td>452</td>
</tr>
<tr>
<td>Inverse score tuning</td>
<td>50</td>
<td>172</td>
<td>48</td>
<td>452</td>
</tr>
<tr>
<td>Model 1</td>
<td>373</td>
<td>6075</td>
<td>465</td>
<td>251</td>
</tr>
<tr>
<td>Model 33</td>
<td>463</td>
<td>5505</td>
<td>718</td>
<td>301</td>
</tr>
<tr>
<td>Second 13-week period</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>536</td>
<td>8361</td>
<td>791</td>
<td>278</td>
</tr>
<tr>
<td>Median</td>
<td>55</td>
<td>1006</td>
<td>71</td>
<td>28</td>
</tr>
<tr>
<td>Ensemble</td>
<td>55</td>
<td>1006</td>
<td>71</td>
<td>28</td>
</tr>
<tr>
<td>Sym trim</td>
<td>54</td>
<td>1020</td>
<td>70</td>
<td>28</td>
</tr>
<tr>
<td>Asym exterior trim</td>
<td>53</td>
<td>826</td>
<td>74</td>
<td>30</td>
</tr>
<tr>
<td>Asym interior trim</td>
<td>50</td>
<td>876</td>
<td>66</td>
<td>27</td>
</tr>
<tr>
<td>Previous best</td>
<td>46</td>
<td>587</td>
<td>66</td>
<td>27</td>
</tr>
<tr>
<td>Inverse score</td>
<td>44</td>
<td>593</td>
<td>62</td>
<td>29</td>
</tr>
<tr>
<td>Inverse score tuning</td>
<td>44</td>
<td>593</td>
<td>62</td>
<td>29</td>
</tr>
<tr>
<td>Model 1</td>
<td>373</td>
<td>6075</td>
<td>465</td>
<td>251</td>
</tr>
<tr>
<td>Model 33</td>
<td>463</td>
<td>5505</td>
<td>718</td>
<td>301</td>
</tr>
<tr>
<td>Third 13-week period</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>1148</td>
<td>14161</td>
<td>2156</td>
<td>391</td>
</tr>
<tr>
<td>Median</td>
<td>1168</td>
<td>18290</td>
<td>1956</td>
<td>405</td>
</tr>
<tr>
<td>Ensemble</td>
<td>1203</td>
<td>17663</td>
<td>2066</td>
<td>434</td>
</tr>
<tr>
<td>Sym trim</td>
<td>1216</td>
<td>15104</td>
<td>2290</td>
<td>399</td>
</tr>
<tr>
<td>Asym exterior trim</td>
<td>1325</td>
<td>18193</td>
<td>2413</td>
<td>417</td>
</tr>
<tr>
<td>Asym interior trim</td>
<td>1320</td>
<td>14325</td>
<td>2341</td>
<td>377</td>
</tr>
<tr>
<td>Previous best</td>
<td>1307</td>
<td>21529</td>
<td>1903</td>
<td>671</td>
</tr>
<tr>
<td>Inverse score</td>
<td>1029</td>
<td>12431</td>
<td>1914</td>
<td>377</td>
</tr>
<tr>
<td>Inverse score tuning</td>
<td>838</td>
<td>11675</td>
<td>1343</td>
<td>396</td>
</tr>
<tr>
<td>Model 1</td>
<td>1155</td>
<td>21363</td>
<td>1520</td>
<td>613</td>
</tr>
<tr>
<td>Model 33</td>
<td>1051</td>
<td>14483</td>
<td>1634</td>
<td>530</td>
</tr>
<tr>
<td>Fourth 13-week period</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>1125</td>
<td>7176</td>
<td>2125</td>
<td>769</td>
</tr>
<tr>
<td>Median</td>
<td>595</td>
<td>6391</td>
<td>682</td>
<td>646</td>
</tr>
<tr>
<td>Ensemble</td>
<td>594</td>
<td>6624</td>
<td>668</td>
<td>645</td>
</tr>
<tr>
<td>Sym trim</td>
<td>890</td>
<td>6968</td>
<td>1499</td>
<td>701</td>
</tr>
<tr>
<td>Asym exterior trim</td>
<td>1078</td>
<td>7176</td>
<td>1987</td>
<td>765</td>
</tr>
<tr>
<td>Asym interior trim</td>
<td>1022</td>
<td>7552</td>
<td>1828</td>
<td>739</td>
</tr>
<tr>
<td>Previous best</td>
<td>856</td>
<td>7529</td>
<td>1076</td>
<td>965</td>
</tr>
<tr>
<td>Inverse score</td>
<td>875</td>
<td>7551</td>
<td>1389</td>
<td>726</td>
</tr>
<tr>
<td>Inverse score tuning</td>
<td>782</td>
<td>7768</td>
<td>1055</td>
<td>758</td>
</tr>
<tr>
<td>Model 1</td>
<td>638</td>
<td>5108</td>
<td>828</td>
<td>690</td>
</tr>
<tr>
<td>Model 33</td>
<td>722</td>
<td>5782</td>
<td>1036</td>
<td>686</td>
</tr>
</tbody>
</table>
Table 6 For cumulative mortality, scores for each quarter of our 52-week dataset. The unit of the scores is deaths. For each quarter, boxed numbers indicate the best method in each column.

<table>
<thead>
<tr>
<th>Method</th>
<th>MIS</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All</td>
<td>U.S.</td>
</tr>
<tr>
<td><strong>First 13-week period</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>1568</td>
<td>14137</td>
</tr>
<tr>
<td>Median</td>
<td>1811</td>
<td>14288</td>
</tr>
<tr>
<td>Model 1</td>
<td>1686</td>
<td>12850</td>
</tr>
</tbody>
</table>

| **Second 13-week period**|     |      |      |     |     |     |      |      |     |     |
| Mean                    | 1099| 15757| 1813 | 505 | 116 | 126 | 2541 | 162  | 56  | 18  |
| Median                  | 1296| 17304| 2136 | 653 | 156 | 128 | 2508 | 167  | 59  | 19  |
| Ensemble                | 1314| 17665| 2163 | 658 | 159 | 127 | 2425 | 166  | 59  | 19  |
| Sym trim                | 1331| 23368| 1990 | 570 | 137 | 127 | 2530 | 163  | 57  | 19  |
| Asym exterior trim      | 1444| 26815| 2081 | 613 | 146 | NA  | NA   | NA   | NA  | NA  |
| Asym interior trim      | 937 | 15757| 1338 | 486 | 115 | NA  | NA   | NA   | NA  | NA  |
| Previous best           | 1153| 10366| 2031 | 680 | 207 | 118 | 1467 | 186  | 68  | 20  |
| Inverse score           | 943 | 12051| 1575 | 489 | 110 | 120 | 2313 | 158  | 56  | 18  |
| Inverse score tuning    | 943 | 12051| 1575 | 489 | 110 | 120 | 2313 | 158  | 56  | 18  |
| Model 1                 | 1282| 16376| 2079 | 685 | 192 | 134 | 1576 | 198  | 91  | 28  |
| Model 33                | 1236| 14458| 2081 | 701 | 149 | 104 | 1317 | 165  | 60  | 17  |

| **Third 13-week period**|     |      |      |     |     |     |      |      |     |     |
| Mean                    | 6747| 103950|13424 |690 | 409 | 445 | 9141 | 661  | 115 | 49  |
| Median                  | 7735| 132486|14540 |790 | 538 | 440 | 8696 | 670  | 115 | 49  |
| Ensemble                | 7436| 115628|14588 |797 | 560 | 435 | 8479 | 669  | 115 | 49  |
| Sym trim                | 6952| 107828|13792 |712 | 418 | 452 | 9276 | 675  | 114 | 48  |
| Asym exterior trim      | 6943| 107979|13733 |727 | 426 | NA  | NA   | NA   | NA  | NA  |
| Asym interior trim      | 4374| 70222 |8224  |747 | 278 | NA  | NA   | NA   | NA  | NA  |
| Previous best           | 6833| 104759|13106 |1231| 402 | 456 | 8002 | 721  | 150 | 55  |
| Inverse score           | 6243| 83841 |13147 |689 | 327 | 439 | 8894 | 659  | 114 | 48  |
| Inverse score tuning    | 6101| 84170 |12660 |741 | 311 | 420 | 8058 | 655  | 110 | 47  |
| Model 1                 | 16575|359767 |24754 |3511| 1271|687  |15911|908   |188 | 69  |
| Model 33                | 6794| 101831 |13186 |1000| 607 | 404 | 7123 | 648  | 120 | 49  |

| **Fourth 13-week period**|     |      |      |     |     |     |      |      |     |     |
| Mean                    | 2661| 22735 |4868 |1693 |243 | 235 | 2216 | 479  | 91  | 20  |
| Median                  | 1848| 13909 |3164 |1443 |229 | 138 | 1743 | 221  | 78  | 19  |
| Ensemble                | 1804| 14326 |2999 |1450 |225 | 139 | 1882 | 216  | 79  | 20  |
| Sym trim                | 2625| 22735 |4885 |1594 |214 | 138 | 1708 | 220  | 81  | 19  |
| Asym exterior trim      | 2714| 22735 |5004 |1724 |237 | NA  | NA   | NA   | NA  | NA  |
| Asym interior trim      | 6939| 21280 |18101|1638 |234 | NA  | NA   | NA   | NA  | NA  |
| Previous best           | 2488| 30259 |3625 |1904 |301 | 174 | 2265 | 269  | 108 | 21  |
| Inverse score           | 2656| 24009 |4791 |1692 |228 | 223 | 2173 | 443  | 91  | 20  |
| Inverse score tuning    | 2163| 24744 |3191 |1729 |240 | 155 | 2211 | 228  | 95  | 20  |
| Model 1                 | 3379| 18523 |6319 |2438 |488 | 194 | 2656 | 310  | 102 | 26  |
| Model 33                | 2246| 14769 |4131 |1588 |283 | 179 | 3271 | 250  | 82  | 24  |
Effect of type of model on performance

Tables 7 and 8 compare the accuracy of the interval forecasts from compartmental models and non-compartmental models of incident and cumulative mortalities, respectively. For most combining methods, for the category including all series, the combined forecasts of incident mortality from forecasts of compartmental models were more accurate than the combined forecasts from forecasts of all models (Table 7), while for most combining methods for cumulative mortality, the reverse was true (Table 8).

**Table 7** For 95% interval forecasts of incident mortality, MIS for all models, compartmental models and non-compartmental models. The unit of the scores is deaths. Boxed numbers indicate the best method in each series.

<table>
<thead>
<tr>
<th>Method</th>
<th>All</th>
<th>Comp</th>
<th>Non-comp</th>
<th>All</th>
<th>Comp</th>
<th>Non-comp</th>
<th>All</th>
<th>Comp</th>
<th>Non-comp</th>
<th>All</th>
<th>Comp</th>
<th>Non-comp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>900</td>
<td>763</td>
<td>1095</td>
<td>1631</td>
<td>1111</td>
<td>2123</td>
<td>443</td>
<td>463</td>
<td>477</td>
<td>104</td>
<td>112</td>
<td>112</td>
</tr>
<tr>
<td>Median</td>
<td>778</td>
<td>890</td>
<td>765</td>
<td>1172</td>
<td>1317</td>
<td>1150</td>
<td>435</td>
<td>488</td>
<td>436</td>
<td>110</td>
<td>125</td>
<td>110</td>
</tr>
<tr>
<td>Sym trim</td>
<td>878</td>
<td>872</td>
<td>967</td>
<td>1515</td>
<td>1227</td>
<td>1807</td>
<td>446</td>
<td>504</td>
<td>440</td>
<td>109</td>
<td>120</td>
<td>109</td>
</tr>
<tr>
<td>Asym ext trim</td>
<td>996</td>
<td>877</td>
<td>1117</td>
<td>1716</td>
<td>1215</td>
<td>2179</td>
<td>480</td>
<td>492</td>
<td>504</td>
<td>118</td>
<td>128</td>
<td>125</td>
</tr>
<tr>
<td>Asym int trim</td>
<td>895</td>
<td>666</td>
<td>1220</td>
<td>1625</td>
<td>920</td>
<td>2517</td>
<td>429</td>
<td>428</td>
<td>462</td>
<td>97</td>
<td>98</td>
<td>105</td>
</tr>
<tr>
<td>Envelope</td>
<td>5284</td>
<td>1241</td>
<td>5070</td>
<td>11037</td>
<td>1475</td>
<td>10840</td>
<td>1004</td>
<td>651</td>
<td>909</td>
<td>300</td>
<td>172</td>
<td>279</td>
</tr>
<tr>
<td>Previous best</td>
<td>877</td>
<td>854</td>
<td>1006</td>
<td>1217</td>
<td>1202</td>
<td>1437</td>
<td>637</td>
<td>517</td>
<td>659</td>
<td>135</td>
<td>128</td>
<td>164</td>
</tr>
<tr>
<td>Inv score</td>
<td>770</td>
<td>637</td>
<td>927</td>
<td>1296</td>
<td>935</td>
<td>1686</td>
<td>425</td>
<td>424</td>
<td>451</td>
<td>98</td>
<td>100</td>
<td>107</td>
</tr>
<tr>
<td>Inv score tuning</td>
<td>677</td>
<td>619</td>
<td>812</td>
<td>1000</td>
<td>859</td>
<td>1298</td>
<td>441</td>
<td>428</td>
<td>477</td>
<td>104</td>
<td>103</td>
<td>110</td>
</tr>
</tbody>
</table>

**Table 8** For 95% interval forecasts of cumulative mortality, MIS for all models, compartmental models and non-compartmental models. The unit of the scores is deaths. Boxed numbers indicate the best method in each series.

<table>
<thead>
<tr>
<th>Method</th>
<th>All</th>
<th>Comp</th>
<th>Non-comp</th>
<th>All</th>
<th>Comp</th>
<th>Non-comp</th>
<th>All</th>
<th>Comp</th>
<th>Non-comp</th>
<th>All</th>
<th>Comp</th>
<th>Non-comp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>3529</td>
<td>3763</td>
<td>3593</td>
<td>6761</td>
<td>6577</td>
<td>7290</td>
<td>925</td>
<td>1012</td>
<td>982</td>
<td>256</td>
<td>308</td>
<td>261</td>
</tr>
<tr>
<td>Median</td>
<td>3700</td>
<td>4232</td>
<td>3528</td>
<td>6769</td>
<td>7292</td>
<td>6685</td>
<td>934</td>
<td>1073</td>
<td>927</td>
<td>310</td>
<td>366</td>
<td>289</td>
</tr>
<tr>
<td>Sym trim</td>
<td>3671</td>
<td>4003</td>
<td>3638</td>
<td>6956</td>
<td>6781</td>
<td>7388</td>
<td>924</td>
<td>1072</td>
<td>927</td>
<td>257</td>
<td>315</td>
<td>256</td>
</tr>
<tr>
<td>Asym ext trim</td>
<td>3737</td>
<td>4028</td>
<td>3746</td>
<td>7006</td>
<td>6885</td>
<td>7450</td>
<td>985</td>
<td>1099</td>
<td>1011</td>
<td>271</td>
<td>332</td>
<td>276</td>
</tr>
<tr>
<td>Asym int trim</td>
<td>3883</td>
<td>2663</td>
<td>3900</td>
<td>8623</td>
<td>4636</td>
<td>8795</td>
<td>921</td>
<td>958</td>
<td>929</td>
<td>208</td>
<td>221</td>
<td>231</td>
</tr>
<tr>
<td>Envelope</td>
<td>6545</td>
<td>3448</td>
<td>6036</td>
<td>11304</td>
<td>5740</td>
<td>10836</td>
<td>1968</td>
<td>1415</td>
<td>1642</td>
<td>663</td>
<td>365</td>
<td>609</td>
</tr>
<tr>
<td>Previous best</td>
<td>3531</td>
<td>3601</td>
<td>3718</td>
<td>6371</td>
<td>6662</td>
<td>6653</td>
<td>1235</td>
<td>1304</td>
<td>1296</td>
<td>303</td>
<td>288</td>
<td>343</td>
</tr>
<tr>
<td>Inv score</td>
<td>3296</td>
<td>3389</td>
<td>3480</td>
<td>6561</td>
<td>6313</td>
<td>7237</td>
<td>920</td>
<td>928</td>
<td>985</td>
<td>221</td>
<td>249</td>
<td>235</td>
</tr>
<tr>
<td>Inv score tuning</td>
<td>3100</td>
<td>3143</td>
<td>3368</td>
<td>5918</td>
<td>6062</td>
<td>6389</td>
<td>948</td>
<td>966</td>
<td>987</td>
<td>219</td>
<td>239</td>
<td>243</td>
</tr>
</tbody>
</table>
Performance of methods at the state level

Fig. 5 presents the MIS for the overall leading method (combining with inverse score with tuning), the popular simple benchmark (the mean combination) and the leading model (Model 33) for forecasts at the individual state level, of incident and cumulative mortalities. In each plot, the states are ordered on the x-axis by the number of cumulative deaths at the end of the final week of our dataset (from low to high). Lower scores are better.

![Incident mortality](image1)

![Cumulative mortality](image2)

**Fig. 5** MIS for 95% interval forecasts of incident and cumulative mortality at (a) high, (b) medium and (c) low mortality U.S. locations for the inverse score with tuning combining (black line), mean combining (dashed line) and Model 33 (dotted line).

Fig. 6 shows the corresponding plots for the MAE. Model 33 is shown, across the states, as being more competitive with the inverse score with tuning in point forecasting (Fig. 6) compared to interval forecasting (Fig. 5). The adverse effects of reporting delays of death counts (highlighted in Fig. 1) on forecast performance is apparent for the states of Ohio, Oklahoma, Delaware and West Virginia, with both combining methods and the individual model underperforming for these states. Figs. 5 and 6 also show the mean combination having similar performance to the leading method for most states in the medium and low mortality categories for both interval and point forecasts of both incident and cumulative mortalities. This is not reflected in Tables 1 to 4 due to adverse effects of reporting delays on performance being greatest for the mean, particularly in the states of Ohio (Figs. 5, 7 and 8) and Delaware (Fig. 6). The impact of reporting delays on interval forecasts differed across...
mortalities. In particular, the inverse score with tuning method was less affected than the mean and Model 33 for forecasts of incident mortality in Ohio (a, in Fig.5, Incident mortality) and forecasts of cumulative mortality in Delaware (c, in Fig.6). Also, the performance of the mean was adversely affected for forecasts of cumulative mortality in Delaware (c, in Fig. 6) but not forecasts of incident mortality in this state (c, in Fig. 5, Incident mortality). The adverse effects of reporting delays on the performance of the other simple benchmark, the median, were lower (not shown).

![Incident mortality vs Cumulative mortality graphs](image.png)

**Fig. 6** MAE for point forecasts of incident and cumulative mortality at (a) high, (b) medium and (c) low mortality U.S. locations, for the inverse score with tuning combining (black line), mean combining (dashed line) and Model 33 (dotted line).

As a sensitivity analysis, we evaluated the impact on our results when excluding all the six states, shown in Fig.1, that had notable effects of reporting delays on death counts. The revised results corresponding to those presented in Tables 1 and 2 are shown in Tables S5 and S6, respectively, in the supplementary information. There were improvements in the performance for all methods, and only slight changes in the ranking of methods. The relative performance of the simple combining methods (mean and median) improved, in comparison with inverse score with tuning. With regard to our significance testing results, there were no changes in our conclusions for the methods that were significantly worse than the method with the best mean rank.
Calibration of interval forecasting methods

The calibration of the interval forecasting methods for incident and cumulative mortalities is summarised in Table 9. The table shows that there was a general tendency for underestimation in the widths of the 95% intervals. This is quite common in studies of interval forecasting (see, for example, (39)). When models underestimate the interval width, combining with the asymmetric interior trimmed mean is useful, as it leads to wider than average intervals. It is, therefore, not surprising that this combining method performs well in Table 9. The envelope method also performed well, which is understandable, because this method also leads to wider than average intervals (as shown in Fig 4). However, this method is not particularly appealing, as it seems likely to deliver intervals that are too wide, and this can be seen in Table 9. We should also note that very sizeable overestimation will lead to a result of 100% for the calibration in Table 9, which is quite close to the ideal of 95%, while sizeable underestimation can lead to calibration far from 95%. This highlights a limitation of calibration for evaluating interval forecasts, and supports our far more extensive use of the interval score in this paper. In Table 9, we also note the relatively good performance of the inverse score methods, which were the best performing combining methods overall, when judged in terms of the MIS.

Table 9  Average calibration of the 95% interval forecasts for incident and cumulative mortality in the 39 week out-of-sample period. Boxed numbers indicate the best method in each column. Ideal is 95%.

<table>
<thead>
<tr>
<th>Method</th>
<th>Incident mortality</th>
<th>Cumulative mortality</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All</td>
<td>U.S.</td>
</tr>
<tr>
<td>Mean</td>
<td>87.2</td>
<td>93.4</td>
</tr>
<tr>
<td>Median</td>
<td>85.6</td>
<td>89.3</td>
</tr>
<tr>
<td>Ensemble</td>
<td>85.3</td>
<td>88.6</td>
</tr>
<tr>
<td>Sym trim</td>
<td>86.6</td>
<td>89.3</td>
</tr>
<tr>
<td>Asym exterior trim</td>
<td>81.9</td>
<td>86.6</td>
</tr>
<tr>
<td>Asym interior trim</td>
<td>91.9</td>
<td>94.0</td>
</tr>
<tr>
<td>Envelope</td>
<td>98.7</td>
<td>100.0</td>
</tr>
<tr>
<td>Previous best</td>
<td>83.7</td>
<td>89.3</td>
</tr>
<tr>
<td>Inverse score</td>
<td>90.0</td>
<td>98.7</td>
</tr>
<tr>
<td>Inverse score tuning</td>
<td>89.7</td>
<td>98.7</td>
</tr>
<tr>
<td>Model 1</td>
<td>91.7</td>
<td>98.0</td>
</tr>
<tr>
<td>Model 14</td>
<td>66.4</td>
<td>85.4</td>
</tr>
<tr>
<td>Model 20</td>
<td>67.2</td>
<td>82.9</td>
</tr>
<tr>
<td>Model 21</td>
<td>67.2</td>
<td>82.9</td>
</tr>
<tr>
<td>Model 22</td>
<td>87.4</td>
<td>77.4</td>
</tr>
<tr>
<td>Model 33</td>
<td>83.9</td>
<td>84.8</td>
</tr>
</tbody>
</table>

NA - not applicable
Discussion

We have provided an empirical comparison of combining methods for point and interval forecasts of incident and cumulative mortality due to COVID-19. Our main findings are that weighted combining performed well for both mortalities, and for both interval and point forecasting. Inverse score with tuning was the most accurate method overall. In the most recent quarter, the median combination and ensemble were leading methods for both mortalities and both interval and point forecasts. The median performed better than the mean in point forecasting of both mortalities. The mean was better than the median in interval forecasts of cumulative mortality, while for interval forecasts of incident deaths, neither was notably better than the other. We were able to include only a few individual models in the comparison, and of these, only two were competitive against the leading combining methods. The leading individual model performed well, particularly for forecasts for the high mortality category, and this model was most accurate in point forecasting incident mortality. For most combining methods, overall, the combined incident forecasts from compartmental models were more accurate than the combined forecasts from all models, whilst the reverse was the case for forecasts of cumulative mortality. The adverse effects on performance of updates in death counts, due to reporting delays, was greatest for the mean combination, with different impacts on the predictions of incident and cumulative mortalities, but, overall, the effects on the comparison were minor.

Drawing comparisons with other studies, this research follows our earlier analysis of point and 95% interval forecasts of cumulative COVID-19 mortality from Weeks 18 to 57 (26). Our current results are based on a later and extended period from Weeks 21 to 72. The overall superiority of the inverse score methods and the relative performance of the mean and the median combinations are consistent with the findings from our earlier analysis of forecasting cumulative mortality, although our current analysis shows that, later in the pandemic, the median becomes more dominant than the mean. In contrast with our earlier study, we have considered individual models and the impact of reporting delays on forecast accuracy. We have highlighted some similarities of results of forecasts of incident mortality and cumulative mortality, and several differences, in terms of the relative performance of the mean and median, the performance of individual models, the impacts of model diversity and reporting delays on forecast accuracy. Our conclusion that forecasts from the combining methods are able to outperform forecasts from individual models is consistent with findings in other studies within and outside the field of epidemic prediction (18-22, 40).

The strengths of this study include our consideration of two sets of data, cumulative and incident mortalities, extending our earlier analysis, which considered only cumulative deaths. Another study strength relating to the scope of this study is our comparison of a broad range of forecasting methods involving individual models, simple standard benchmark methods and more complex combinations based on trimming or weighting according to historical accuracy. A further study strength relates to our choice of data. The dataset of forecasts, produced for multiple locations from multiple models, presented an opportunity to study the wisdom of the crowd. The Hub provided the necessary conditions for the crowd being ‘wise’ (14) and without distortion, such as by social pressure (41) or restrictions against forecasting teams applying their own judgement (23). These conditions include independent contributors, diversity of opinions, and a trustworthy central convener to collate the information provided (14).
Study limitations include the retrospective design, being based on the most recent version of the ‘truth data’ for all the weeks at the time of analysis, instead of the numbers of COVID-19 deaths that were reported at the time the forecasts were submitted. We have shown that there were several states for which there were notable effects of updates in death counts, due to reporting delays, and this adversely affected the accuracy of the forecasts of all the combining methods and models. However, this issue only had a minor effect on the relative performances of the methods, and did not alter our overall conclusions. Our reported findings are limited to U.S. data and the forecasts from the COVID-19 Forecast Hub, and so it is possible that different results may arise when applying the combining methods to forecasts from a different set of models, or using other data, such as forecasts for other locations, or predictions of COVID-19 cases or hospitalisations. These are interesting potential avenues for future research. The forecasts in our dataset were produced weekly for 1 to 4 weeks ahead horizons, and we acknowledge that conclusions could be different for different time-scales. Our ability to detect statistical differences was limited by the small sample sizes, with only 17 locations in each category, missing data and a relatively short out-of-sample period.

This research has important policy implications as forecasts from models have been placed at the forefront of public health decision making during the COVID-19 crisis. The reliance of governments on forecasts from COVID-19 models has brought these models under increased scrutiny. The limitations of the models and the need to appreciate their limits have been discussed extensively (2, 8, 9, 11, 42). It is suggested that relying on modelling alone leads to “missteps and blind spots”, and that the best approach to support public policy decision making would involve a triangulation of insights from modelling with other information, such as analyses of international case studies and previous outbreaks, policy documents, and discussions with frontline staff (42). It is essential that modelling offers the most accurate forecasts. The associated uncertainty should be accurately represented in forecasts from models. The 95% prediction intervals present a range of possible outcomes, which can be used to support situational awareness (43). Given the benefits of combining, the most accurate probabilistic forecasts will most often be based on multiple models, rather than an individual model, as illustrated by the results of our study. Although individual models can sometimes be more accurate than combined methods, relying on forecasts from combined methods provides a more risk-averse approach, as the best model will not be clear until records of historical accuracy are available, and also the best performing model will typically change over time. In particular, our finding that the performance of the average (mean combination) was substantially better than the average performance of the individual models suggests that, at the start of an epidemic, when it is not clear which model has the best performance, the statistical expectation is that the average method will score better than a model chosen randomly, or chosen on the basis of no prior history. This study follows on from our previous analysis to provide further confirmation that the weighted methods provided the greatest accuracy, but the performance of the mean and median combination were often reasonable. An obvious advantage of these benchmark methods is that they are simple, transparent and pragmatic approaches to combining forecasts that do not rely on the need to optimise a parameter or collect data on historical accuracy. This is advantageous early in a pandemic, and also later on, if records of historical accuracy are uneven for the individual forecasting models.
To conclude, we recommend that for COVID-19 models to play the most effective role in supporting health policy decision making, probabilistic forecasts should be harnessed from multiple models. In our study of mortality data, the intuitive weighted forecasting methods were the most accurate overall, but that did not rule out considering the more pragmatic simple combining methods. We identified that the relative performance of the different combining methods depends on several factors including the type of data, type of forecast and timing. Future research could focus on seeking further clarification of the relative performance of the different methods, by studying combined forecasts for other types of data, such as COVID-19 infections and hospitalisations, combined forecasts in other locations and for diseases beyond the COVID-19 pandemic. Another possible avenue for further research would be to investigate the impact of incorporating combined forecasts into health policy decision making.

**Funding**
No funding was received for conducting this study.

**Competing interests**
The authors have no relevant financial or non-financial interests to disclose.

**Ethical approval**
No ethical approval was required for this study.

**Data sharing**
This study is based on publically available data from the COVID-19 Forecast Hub.
https://covid19forecasthub.org/

**Acknowledgements**
We are very grateful to the forecasting groups who generously made their forecasts available on the COVID-19 Forecast Hub, Nicholas Reich and his team, for acting as curators for the Hub, and Nia Roberts for clarifying our understanding of the license terms for the forecast data.

**References**


42. Sridhar D, Majumder MS. Modelling the pandemic. BMJ. 2020;369:m1567.

Fig S1 Numbers of weekly incident COVID-19 deaths in U.S. locations between weeks ending 23 May 2020 and 15 May 2021
Fig S2 Numbers of weekly cumulative COVID-19 deaths in U.S. locations between weeks ending 23 May 2020 and 15 May 2021.
**Table S1**  Frequency of being one of the top three methods in the columns for the scores in Tables 1 to 4

<table>
<thead>
<tr>
<th>Method</th>
<th>Total</th>
<th>Incident</th>
<th>Cumulative</th>
<th>Interval</th>
<th>Point</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Median</td>
<td>6</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Ensemble</td>
<td>6</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>Symm trim</td>
<td>6</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Asym exterior trim</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>NA</td>
</tr>
<tr>
<td>Asym interior trim</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>NA</td>
</tr>
<tr>
<td>Envelope</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>NA</td>
</tr>
<tr>
<td>Previous best</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Inverse score</td>
<td>9</td>
<td>4</td>
<td>5</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>Inverse score tuning</td>
<td>16</td>
<td>7</td>
<td>9</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Model 1</td>
<td>6</td>
<td>6</td>
<td>NA</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Model 14</td>
<td>0</td>
<td>0</td>
<td>NA</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Model 20</td>
<td>0</td>
<td>0</td>
<td>NA</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Model 21</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Model 22</td>
<td>0</td>
<td>0</td>
<td>NA</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Model 33</td>
<td>11</td>
<td>7</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
</tbody>
</table>

NA – not applicable

**Table S2**  Frequency of being one of the top three methods in the columns for the mean ranks in Tables 1 to 4

<table>
<thead>
<tr>
<th>Method</th>
<th>Total</th>
<th>Incident</th>
<th>Cumulative</th>
<th>Interval</th>
<th>Point</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Median</td>
<td>7</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Ensemble</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>Symm trim</td>
<td>5</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Asym exterior trim</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>NA</td>
</tr>
<tr>
<td>Asym interior trim</td>
<td>8</td>
<td>3</td>
<td>5</td>
<td>8</td>
<td>NA</td>
</tr>
<tr>
<td>Envelope</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>NA</td>
</tr>
<tr>
<td>Previous best</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Inverse score</td>
<td>12</td>
<td>6</td>
<td>6</td>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td>Inverse score tuning</td>
<td>13</td>
<td>5</td>
<td>8</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>Model 1</td>
<td>0</td>
<td>0</td>
<td>NA</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Model 14</td>
<td>3</td>
<td>3</td>
<td>NA</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Model 20</td>
<td>0</td>
<td>0</td>
<td>NA</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Model 21</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Model 22</td>
<td>0</td>
<td>0</td>
<td>NA</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Model 33</td>
<td>7</td>
<td>5</td>
<td>2</td>
<td>2</td>
<td>5</td>
</tr>
</tbody>
</table>

NA – not applicable
Table S3 For incident mortality, scores in the 39 week out-of-sample period for each of the four lead times. The unit of the scores is deaths. Lower values are better. For each horizon, boxed numbers indicate the best method for each of the following four categories of series: all, U.S., high, medium and low.

<table>
<thead>
<tr>
<th>Method</th>
<th>MIS</th>
<th>All</th>
<th>U.S.</th>
<th>High</th>
<th>Med</th>
<th>Low</th>
<th>All</th>
<th>U.S.</th>
<th>High</th>
<th>Med</th>
<th>Low</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1 week ahead</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>687</td>
<td>6877</td>
<td>1224</td>
<td>386</td>
<td>88</td>
<td>68</td>
<td>1202</td>
<td>92</td>
<td>34</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td>618</td>
<td>6879</td>
<td>914</td>
<td>370</td>
<td>95</td>
<td>64</td>
<td>1152</td>
<td>86</td>
<td>33</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>Ensemble</td>
<td>621</td>
<td>8541</td>
<td>927</td>
<td>374</td>
<td>95</td>
<td>65</td>
<td>1181</td>
<td>85</td>
<td>33</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>Sym trim</td>
<td>638</td>
<td>7222</td>
<td>1075</td>
<td>364*</td>
<td>88†</td>
<td>65*</td>
<td>1152</td>
<td>87</td>
<td>32*</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>Asym interior trim</td>
<td>647</td>
<td>6999</td>
<td>1125</td>
<td>361</td>
<td>81*</td>
<td>68</td>
<td>1223</td>
<td>91</td>
<td>33</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>Inverse score</td>
<td>594*</td>
<td>6506</td>
<td>988*</td>
<td>364*</td>
<td>84*</td>
<td>66*</td>
<td>1159</td>
<td>89</td>
<td>33</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>Inverse score tuning</td>
<td>533*</td>
<td>6245</td>
<td>807</td>
<td>370</td>
<td>87†</td>
<td>62*</td>
<td>1032</td>
<td>85</td>
<td>34*</td>
<td>11*</td>
<td></td>
</tr>
<tr>
<td>Model 21</td>
<td>527*</td>
<td>6005</td>
<td>770†</td>
<td>394</td>
<td>94</td>
<td>62</td>
<td>1039</td>
<td>84</td>
<td>33</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Model 33</td>
<td>710</td>
<td>8352</td>
<td>1121</td>
<td>449</td>
<td>112</td>
<td>66</td>
<td>1081</td>
<td>93</td>
<td>34</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td><strong>2 weeks ahead</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>892</td>
<td>9345</td>
<td>1644</td>
<td>436</td>
<td>97</td>
<td>99</td>
<td>1486</td>
<td>162</td>
<td>40</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td>705</td>
<td>9795</td>
<td>1072</td>
<td>407</td>
<td>100</td>
<td>76</td>
<td>1385</td>
<td>102</td>
<td>38*</td>
<td>12*</td>
<td></td>
</tr>
<tr>
<td>Ensemble</td>
<td>713</td>
<td>9739</td>
<td>1087</td>
<td>418</td>
<td>101</td>
<td>77</td>
<td>1426</td>
<td>102</td>
<td>38*</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>Sym trim</td>
<td>841</td>
<td>9761</td>
<td>1477</td>
<td>424</td>
<td>98</td>
<td>94*</td>
<td>1390</td>
<td>156</td>
<td>38*</td>
<td>12*</td>
<td></td>
</tr>
<tr>
<td>Asym interior trim</td>
<td>913</td>
<td>9322</td>
<td>1743</td>
<td>412</td>
<td>91</td>
<td>99</td>
<td>1514</td>
<td>161</td>
<td>39</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>Inverse score</td>
<td>742*</td>
<td>8069</td>
<td>1285*</td>
<td>418</td>
<td>91*</td>
<td>96</td>
<td>1385*</td>
<td>161</td>
<td>39</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>Inverse score tuning</td>
<td>635*</td>
<td>7693</td>
<td>958</td>
<td>434</td>
<td>96</td>
<td>75</td>
<td>1162*</td>
<td>108</td>
<td>40</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>Model 21</td>
<td>645</td>
<td>8660</td>
<td>888</td>
<td>473</td>
<td>102</td>
<td>74</td>
<td>1237</td>
<td>103</td>
<td>40</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>Model 33</td>
<td>720</td>
<td>9218</td>
<td>1088</td>
<td>454</td>
<td>120</td>
<td>73</td>
<td>1155</td>
<td>105</td>
<td>38</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td><strong>3 weeks ahead</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>979</td>
<td>9660</td>
<td>1863</td>
<td>459</td>
<td>105</td>
<td>107</td>
<td>1809</td>
<td>163</td>
<td>46</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td>799</td>
<td>12427</td>
<td>1160</td>
<td>446</td>
<td>107</td>
<td>92</td>
<td>1726</td>
<td>124</td>
<td>43*</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>Ensemble</td>
<td>813</td>
<td>11915</td>
<td>1217</td>
<td>459</td>
<td>108</td>
<td>92</td>
<td>1708</td>
<td>124</td>
<td>44</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>Sym trim</td>
<td>953</td>
<td>11156</td>
<td>1694</td>
<td>456</td>
<td>110</td>
<td>103</td>
<td>1722</td>
<td>157</td>
<td>44</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>Asym interior trim</td>
<td>942</td>
<td>9921</td>
<td>1751</td>
<td>447</td>
<td>100</td>
<td>107</td>
<td>1825</td>
<td>162</td>
<td>45</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>Inverse score</td>
<td>815*</td>
<td>8952</td>
<td>1432*</td>
<td>436</td>
<td>100*</td>
<td>103*</td>
<td>1666*</td>
<td>159</td>
<td>45</td>
<td>14*</td>
<td></td>
</tr>
<tr>
<td>Inverse score tuning</td>
<td>686</td>
<td>8912</td>
<td>1018</td>
<td>452</td>
<td>105</td>
<td>88</td>
<td>1451</td>
<td>124</td>
<td>46</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>Model 21</td>
<td>806</td>
<td>13889</td>
<td>988</td>
<td>548</td>
<td>113</td>
<td>90</td>
<td>1632</td>
<td>119</td>
<td>46</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>Model 33</td>
<td>698</td>
<td>6844</td>
<td>1097</td>
<td>495</td>
<td>140</td>
<td>80</td>
<td>1191</td>
<td>119</td>
<td>43</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td><strong>4 weeks ahead</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>1044</td>
<td>13315</td>
<td>1792</td>
<td>491</td>
<td>126</td>
<td>121</td>
<td>2347</td>
<td>165</td>
<td>52</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td>989</td>
<td>14055</td>
<td>1543</td>
<td>518</td>
<td>138</td>
<td>112</td>
<td>2263</td>
<td>143</td>
<td>49</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>Ensemble</td>
<td>1029</td>
<td>14509</td>
<td>1611</td>
<td>539</td>
<td>143</td>
<td>111</td>
<td>2204</td>
<td>145</td>
<td>50</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>Sym trim</td>
<td>1077</td>
<td>13641</td>
<td>1814</td>
<td>542</td>
<td>138</td>
<td>118</td>
<td>2263</td>
<td>163</td>
<td>51</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>Asym interior trim</td>
<td>1076</td>
<td>13634</td>
<td>1879</td>
<td>495</td>
<td>116*</td>
<td>121</td>
<td>2353</td>
<td>165</td>
<td>52</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>Inverse score</td>
<td>929*</td>
<td>12993</td>
<td>1478*</td>
<td>481</td>
<td>117</td>
<td>116*</td>
<td>2130*</td>
<td>162</td>
<td>51</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>Inverse score tuning</td>
<td>854*</td>
<td>12906</td>
<td>1216</td>
<td>508</td>
<td>128</td>
<td>104*</td>
<td>1777</td>
<td>145</td>
<td>53</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>Model 21</td>
<td>939</td>
<td>16141</td>
<td>1145</td>
<td>632</td>
<td>146</td>
<td>106</td>
<td>1931</td>
<td>141</td>
<td>55</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>Model 33</td>
<td>863</td>
<td>10312</td>
<td>1249</td>
<td>598</td>
<td>185</td>
<td>99</td>
<td>1680</td>
<td>137</td>
<td>51</td>
<td>17</td>
<td></td>
</tr>
</tbody>
</table>

* and † indicate a score significantly lower than the mean combination and median combination, respectively, at the 5% significance level.
Table S4 For cumulative mortality, scores in the 39 week out-of-sample period for each of the four lead times. The unit of the scores is deaths. For each horizon, boxed numbers indicate the best method for each of the following four categories of series: all series, U.S., high, medium and low.

<table>
<thead>
<tr>
<th>Method</th>
<th>MIS</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All</td>
<td>U.S.</td>
</tr>
<tr>
<td>1 week ahead</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>3319</td>
<td>62108</td>
</tr>
<tr>
<td>Median</td>
<td>3123</td>
<td>57217</td>
</tr>
<tr>
<td>Ensemble</td>
<td>3077</td>
<td>54979</td>
</tr>
<tr>
<td>Sym trim</td>
<td>3277</td>
<td>63747</td>
</tr>
<tr>
<td>Asym interior trim</td>
<td>2204††</td>
<td>33641††</td>
</tr>
<tr>
<td>Previous best</td>
<td>2559††</td>
<td>31436†</td>
</tr>
<tr>
<td>Inverse score</td>
<td>3195</td>
<td>55789*</td>
</tr>
<tr>
<td>Inverse score tuning</td>
<td>2736††</td>
<td>39628†</td>
</tr>
<tr>
<td>Model 33</td>
<td>3271</td>
<td>62685</td>
</tr>
<tr>
<td>2 weeks ahead</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>3045</td>
<td>38620</td>
</tr>
<tr>
<td>Median</td>
<td>3303</td>
<td>47609</td>
</tr>
<tr>
<td>Ensemble</td>
<td>3231</td>
<td>44870</td>
</tr>
<tr>
<td>Sym trim</td>
<td>3105</td>
<td>39643</td>
</tr>
<tr>
<td>Asym interior trim</td>
<td>3491</td>
<td>32197</td>
</tr>
<tr>
<td>Previous best</td>
<td>3063</td>
<td>40956</td>
</tr>
<tr>
<td>Inverse score</td>
<td>2842††</td>
<td>31725</td>
</tr>
<tr>
<td>Inverse score tuning</td>
<td>2756††</td>
<td>32620</td>
</tr>
<tr>
<td>Model 33</td>
<td>3332</td>
<td>48704</td>
</tr>
<tr>
<td>3 weeks ahead</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>3594</td>
<td>43109</td>
</tr>
<tr>
<td>Median</td>
<td>3795</td>
<td>52918</td>
</tr>
<tr>
<td>Ensemble</td>
<td>3655</td>
<td>46027</td>
</tr>
<tr>
<td>Sym trim</td>
<td>3777</td>
<td>48037</td>
</tr>
<tr>
<td>Asym interior trim</td>
<td>4503</td>
<td>36011</td>
</tr>
<tr>
<td>Previous best</td>
<td>3766</td>
<td>53485</td>
</tr>
<tr>
<td>Inverse score</td>
<td>3349††</td>
<td>35893</td>
</tr>
<tr>
<td>Inverse score tuning</td>
<td>3177††</td>
<td>39846</td>
</tr>
<tr>
<td>Model 33</td>
<td>3442</td>
<td>35859</td>
</tr>
<tr>
<td>4 weeks ahead</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>4158</td>
<td>50151</td>
</tr>
<tr>
<td>Median</td>
<td>4580</td>
<td>66996</td>
</tr>
<tr>
<td>Ensemble</td>
<td>4387</td>
<td>56047</td>
</tr>
<tr>
<td>Sym trim</td>
<td>4526</td>
<td>58825</td>
</tr>
<tr>
<td>Asym interior trim</td>
<td>5335</td>
<td>42720</td>
</tr>
<tr>
<td>Previous best</td>
<td>4734</td>
<td>70708</td>
</tr>
<tr>
<td>Inverse score</td>
<td>3799†</td>
<td>38510</td>
</tr>
<tr>
<td>Inverse score tuning</td>
<td>3731</td>
<td>50936</td>
</tr>
<tr>
<td>Model 33</td>
<td>3781</td>
<td>30433</td>
</tr>
</tbody>
</table>

* and † indicate a score significantly lower than the mean combination and median combination, respectively, at the 5% significance level.
### Table S5  Sensitivity analysis for incident mortality, MIS and mean ranks for 95% interval forecasts for the 39 week out-of-sample period excluding the six U.S. states with notable effects of reporting delays on COVID-19 death counts. Boxed numbers indicate the best method in each column.

<table>
<thead>
<tr>
<th>Method</th>
<th>MIS</th>
<th>Mean rank</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All</td>
<td>U.S.</td>
</tr>
<tr>
<td>Mean</td>
<td>789</td>
<td>9799</td>
</tr>
<tr>
<td>Median</td>
<td>723</td>
<td>11239</td>
</tr>
<tr>
<td>Ensemble</td>
<td>737</td>
<td>11176</td>
</tr>
<tr>
<td>Sym trim</td>
<td>758</td>
<td>10445</td>
</tr>
<tr>
<td>Asym exterior trim</td>
<td>888</td>
<td>12447</td>
</tr>
<tr>
<td>Asym interior trim</td>
<td>816</td>
<td>9969</td>
</tr>
<tr>
<td>Envelope</td>
<td>5042</td>
<td>64944</td>
</tr>
<tr>
<td>Previous best</td>
<td>864</td>
<td>11779</td>
</tr>
<tr>
<td>Inverse score tuning</td>
<td>686</td>
<td>9130</td>
</tr>
<tr>
<td>Intervals score tuning</td>
<td>651</td>
<td>8939</td>
</tr>
<tr>
<td>Model 22</td>
<td>993</td>
<td>19145</td>
</tr>
<tr>
<td>Model 21</td>
<td>724</td>
<td>11174</td>
</tr>
<tr>
<td>Model 14</td>
<td>1961</td>
<td>30502</td>
</tr>
<tr>
<td>Model 20</td>
<td>2005</td>
<td>33948</td>
</tr>
<tr>
<td>Model 21</td>
<td>2056</td>
<td>35399</td>
</tr>
<tr>
<td>Model 33</td>
<td>705</td>
<td>8681</td>
</tr>
<tr>
<td>Mean MIS of models</td>
<td>2208</td>
<td>37870</td>
</tr>
</tbody>
</table>

### Table S6  Sensitivity analysis for cumulative mortality, MIS and mean ranks for 95% interval forecasts for the 39 week out-of-sample period excluding the six U.S. states with notable effects of reporting delays on COVID-19 death counts. Boxed numbers indicate the best method in each column.

<table>
<thead>
<tr>
<th>Method</th>
<th>MIS</th>
<th>Mean rank</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All</td>
<td>U.S.</td>
</tr>
<tr>
<td>Mean</td>
<td>2003</td>
<td>48497</td>
</tr>
<tr>
<td>Median</td>
<td>2257</td>
<td>56185</td>
</tr>
<tr>
<td>Ensemble</td>
<td>2145</td>
<td>50481</td>
</tr>
<tr>
<td>Sym trim</td>
<td>2141</td>
<td>52563</td>
</tr>
<tr>
<td>Asym exterior trim</td>
<td>2219</td>
<td>53839</td>
</tr>
<tr>
<td>Asym interior trim</td>
<td>1741</td>
<td>36142</td>
</tr>
<tr>
<td>Envelope</td>
<td>4738</td>
<td>103456</td>
</tr>
<tr>
<td>Previous best</td>
<td>2252</td>
<td>49146</td>
</tr>
<tr>
<td>Inverse score tuning</td>
<td>1778</td>
<td>40479</td>
</tr>
<tr>
<td>Model 21</td>
<td>6015</td>
<td>136492</td>
</tr>
<tr>
<td>Model 33</td>
<td>2074</td>
<td>44297</td>
</tr>
<tr>
<td>Mean MIS of models</td>
<td>5229</td>
<td>114911</td>
</tr>
</tbody>
</table>
## Appendix  Forecasting models

<table>
<thead>
<tr>
<th>Contributors</th>
<th>Short model name</th>
<th>Model description*</th>
<th>Access and licencing information</th>
<th>Citations</th>
</tr>
</thead>
<tbody>
<tr>
<td>COMPARTMENTAL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pei S, Yamana T, Kandula S, Yang W, Galanti M, Shaman J</td>
<td>CU-select</td>
<td>Metapopulation county-level SEIR model for projecting future COVID-19 incidence and deaths. This forecast is the scenario we believe to be most plausible given the current setting.</td>
<td><a href="https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/CU-select">https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/CU-select</a></td>
<td><a href="https://doi.org/10.1101/2020.03.21.20040303">https://doi.org/10.1101/2020.03.21.20040303</a> <a href="https://www.medrxiv.org/content/10.1101/2020.05.04.20090670v2">https://www.medrxiv.org/content/10.1101/2020.05.04.20090670v2</a></td>
</tr>
<tr>
<td>Model Name</td>
<td>Method Description</td>
<td>Paper and Code Link</td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------------------</td>
<td>------------------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kinsey M, Tallaksen K, Obrecht RF, Asher L, Costello C, Kelbaugh M, Wilson S</td>
<td>Metapopulation model using public mobility data. Local parameters (case reporting rates, doubling times, etc) are estimated using data from CSSE and CDC scenario 5. Primary output is case incidence.</td>
<td><a href="https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/JHUAPL_Bucky">https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/JHUAPL_Bucky</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Authors</td>
<td>Model</td>
<td>Description</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------</td>
<td>-------</td>
<td>-------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gao Z, Li C, Zheng S, Bian J, Xie X, Liu T-Y</td>
<td>NotreDame-Mobility</td>
<td>Ensemble of nine models that are identical except that they are driven by different mobility indices from Apple and Google. The model underlying each is a deterministic, SEIR-like model.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Turtle J, Ben-Nun M, Riley P</td>
<td>PSI-DRAFT</td>
<td>A stochastic/deterministic, single-population SEIRX model that stratifies by both age distribution and disease severity and includes generic intervention fitting.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shi Y, Shah T, Ban X</td>
<td>RPI-UW-Mob_Collision</td>
<td>A mobility-informed simplified SIR model motivated by collision theory.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Snyder TL, Wilson DD</td>
<td>SWC-TerminusCM</td>
<td>Mechanistic compartmental model using disease parameter estimates from literature. It uses Bayesian inference to predict the most likely model parameters.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Authors</td>
<td>Institute or Source</td>
<td>Description</td>
<td>Website</td>
<td></td>
</tr>
<tr>
<td>--------------------</td>
<td>-------------------------------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>Gu Q, Xu P, Chen J, Wang L, Zou D, Zhang W</td>
<td>UCLA-SuEIR</td>
<td>Variant of the SEIR model considering both untested and unreported cases. The model considers reopening and assumes susceptible population will increase after the reopen.</td>
<td><a href="https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/UCLA-SuEIR">https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/UCLA-SuEIR</a></td>
<td></td>
</tr>
<tr>
<td>Chen YQ, Zhao Y, Guo L</td>
<td>UCM-MESALab-FoGSEIR</td>
<td>FoGSEIR model is a modification of integer order SEIR model considering fractional integrals. The model considers the age structure and reopening intervention to minimize infections and deaths.</td>
<td><a href="https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/UCM_MESALab-FoGSEIR">https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/UCM_MESALab-FoGSEIR</a></td>
<td></td>
</tr>
<tr>
<td>Mayo ML, Rowland MA, Parno MD, Detwiller ID, Farthing MW, England WP George GE</td>
<td>USACE-ERDC_SEIR</td>
<td>The ERDC SEIR model makes predictions of several variables (e.g., reported new/cumulative cases per day, etc.). Model parameters are estimated using historical data using Bayesian inference.</td>
<td><a href="https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/USACE-ERDC_SEIR">https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/USACE-ERDC_SEIR</a></td>
<td></td>
</tr>
<tr>
<td>Gu Y</td>
<td>YYG-ParamSearch</td>
<td>Based on the SEIR model with hyperparameter optimization to make daily projections regarding COVID-19 infections and deaths in 50 US states. The model accounts for state reopenings and its effects on infections and deaths.</td>
<td><a href="https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/YYG-ParamSearch">https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/YYG-ParamSearch</a></td>
<td></td>
</tr>
</tbody>
</table>

NON-COMPARTMENTAL

<table>
<thead>
<tr>
<th>Authors</th>
<th>Institute or Source</th>
<th>Description</th>
<th>Website</th>
</tr>
</thead>
<tbody>
<tr>
<td>Team</td>
<td>Model</td>
<td>Description</td>
<td>GitHub Repository</td>
</tr>
<tr>
<td>------</td>
<td>-------</td>
<td>-------------</td>
<td>-------------------</td>
</tr>
</tbody>
</table>

§ CMU-Timeseries is a basic AR-type time series model fit using case counts and deaths as features.
| Murry C and the IHME-CurveFitTeam | IHME-CurveFit | Non-linear mixed effects curve-fitting. This model makes predictions about the future that are dependent on the assumption that current interventions continue. | https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/IHME-CurveFit |
| Wang L, Wang G, Gao L, Li X, Yu S, Kim M, Wang Y, Gu Z. | IowaStateLW-STEM | A nonparametric space-time disease transmission model. The projections assume that the data used is reliable, the future will continue to follow the current pattern, and current interventions will remain the same till the end of forecasting period. | https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/IowaStateLW-STEM |
| Karlem D | Karlen-pypm | Discrete-time difference equations with long periods of constant transmission rate | https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/Karlen-pypm |
| Espana G, Oidtman R, Cavany S, Costello A, Wieler A, | NotreDame-FRED | Agent-based model developed for influenza with parameters modified to | https://github.com/reichlab/covid19-forecast-
<table>
<thead>
<tr>
<th>Authors</th>
<th>Model/Method</th>
<th>Details</th>
<th>GitHub Repository</th>
</tr>
</thead>
<tbody>
<tr>
<td>McConnell S, Donaldson B</td>
<td>SteveMcConnell_CovidComplete</td>
<td>A near-term fatality prediction model that calculates and uses fatality trends at the national and state level, trends in positive virus tests and total virus tests, and age-related demographics for state forecasts. Model forecasts are based on predicting near-term deaths from recent positive virus tests.</td>
<td><a href="https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/SteveMcConnell-CovidComplete">https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/SteveMcConnell-CovidComplete</a></td>
</tr>
<tr>
<td>Bieggel H, Lega J</td>
<td>UA-EpiCovDA</td>
<td>SIR mechanistic model with data assimilation. EpiCovDA is an extension of the EpiGro model. Model parameters are fit to Covid-19 data using a variational data assimilation method. A prior distribution of the parameters is estimated by fitting an SIR Incidence-Cumulative Cases curve to data from states that had at least 1000 cases by 04/01/2020.</td>
<td><a href="https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/UA-EpiCovDA">https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/UA-EpiCovDA</a></td>
</tr>
<tr>
<td>Jin X, Wang Y-X, Yan X</td>
<td>UCSB-ACTS</td>
<td>This data-driven machine learning model makes predictions by referring to other regions with similar growth patterns and assuming the similar development will take place in the current region.</td>
<td><a href="https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/UCSB-ACTS">https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/UCSB-ACTS</a></td>
</tr>
<tr>
<td>Model Name</td>
<td>GitHub Link</td>
<td>Details</td>
<td></td>
</tr>
<tr>
<td>------------</td>
<td>-------------</td>
<td>---------</td>
<td></td>
</tr>
<tr>
<td>Srivastava A, Prasanna VK, Xu FT</td>
<td><a href="https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/USC-SI_kJalpha">https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/USC-SI_kJalpha</a></td>
<td>A heterogeneous infection rate model with human mobility for epidemic modeling. Our model adapts to changing trends and provide predictions of confirmed cases and deaths. We build a random forest, based on the output of USC_SIKJalpha model along with the data on the cumulative case/death, weekly increase, and previous increase. We then sample trees to generate quantile forecasts.</td>
<td></td>
</tr>
<tr>
<td>Woody S, et al. at the University of Texas</td>
<td><a href="https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/UT-Mobility">https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/UT-Mobility</a></td>
<td>This model makes predictions assuming that social distancing patterns, as measured by anonymized mobile-phone GPS traces, remain constant in the future. Only models <em>first-wave deaths</em>.</td>
<td></td>
</tr>
</tbody>
</table>

* Based on information recorded on the COVID19 Hub with citations as recorded on 18/5/21; † Only provided forecasts of numbers of cumulative COVID-19 deaths; § Only provided forecasts of numbers of incident COVID-19 deaths.