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

Kathryn S. Taylor[1,*] James W Taylor[2]

* Corresponding author kathryn.taylor@phc.ox.ac.uk

Affiliations

[1]Nuffield Department of Primary Care Health Sciences, University of Oxford

[2] Saïd 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.

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.

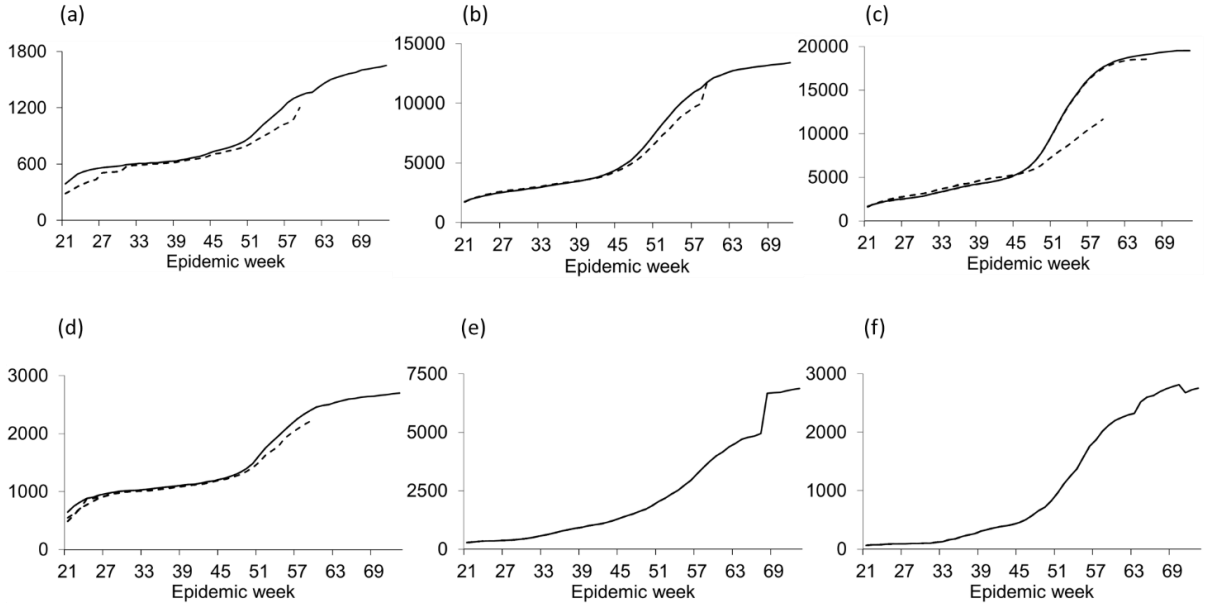

Fig. 1 Numbers of reported deaths in states where there were notable effects of reporting delays: (a) Delaware, (b) Indiana, (c) Ohio, (d) Rhode Island, (e) Oklahoma and (f) West Virginia. Updates with backdating (a-d) are shown as deviations (dashed lines) from the most recent plot of cumulative COVID-19 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.

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.

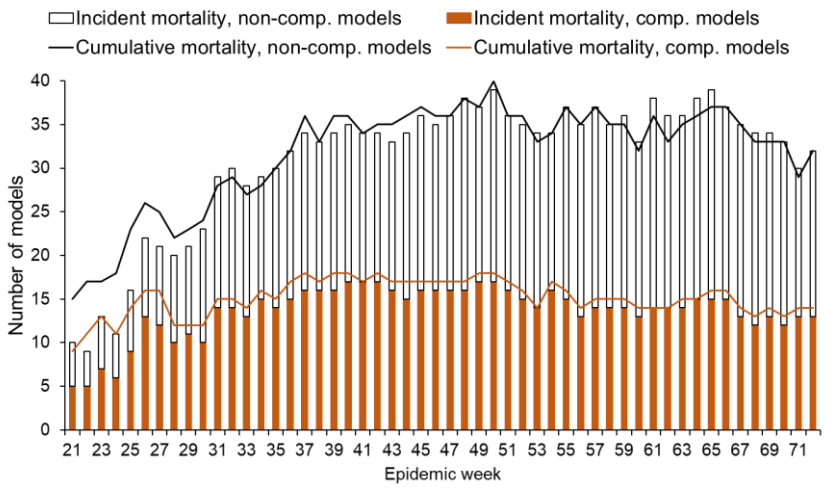

Fig. 2 Number and types of models at each forecast origin for each mortality dataset

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.

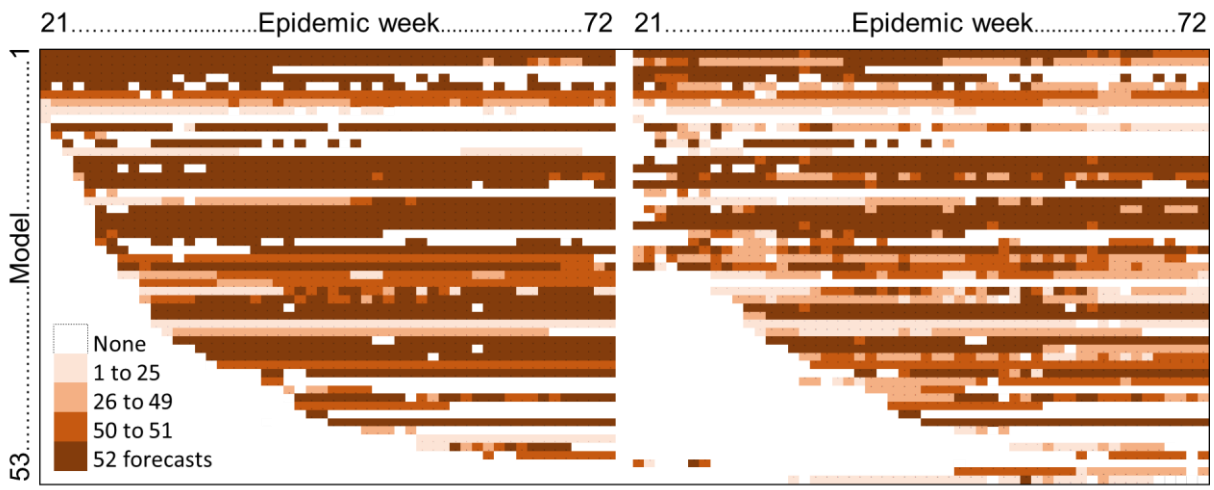

Fig. 3 Extent of missing data for forecasts of incident mortality (left) and cumulative mortality (right) from Models 1 to 53 for forecast origins between Epidemic Weeks 21 and 72

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):

$$S_{\alpha}^{INT}\left(l_t, u_t, y_t\right) = \left(u_t - l_t\right) + \frac{2}{\alpha} I\left\{y_t \le l_t\right\}\left(l_t - y_t\right) + \frac{2}{\alpha} I\left\{y_t \ge u_t\right\}\left(y_t - u_t\right)$$

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, α=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 α (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.

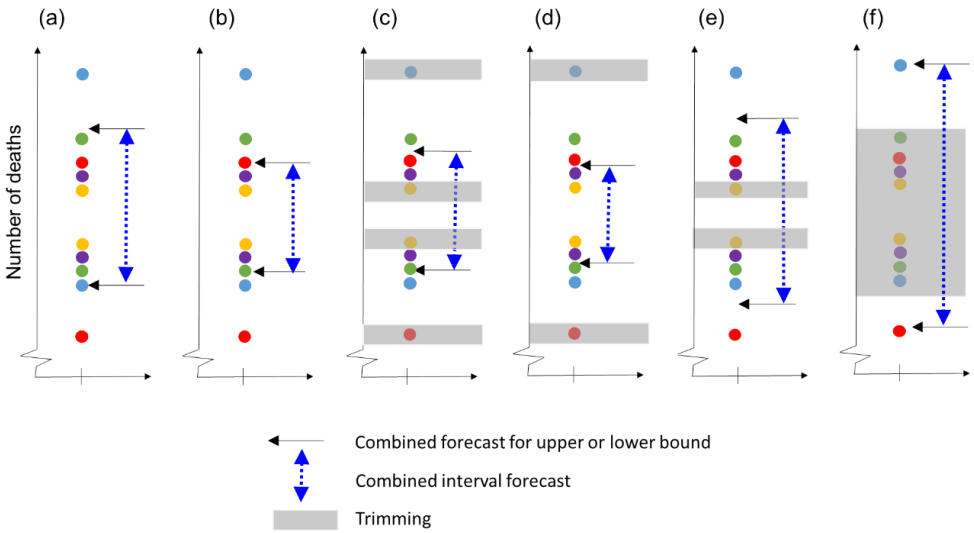

Fig. 4 Illustration of non-score-based combining methods of interval forecasts from five individual models, showing (a) mean combination, (b) median combination, (c) symmetric trimmed mean, (d) asymmetric exterior trimmed mean, (e) asymmetric interior trimmed mean and (f) envelope. Each interval forecast is represented by a different coloured pair of dots.

More novel methods of combining involve trimming (excluding) a particular percentage β 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 β 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 β 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):

$$w_{it} = \frac{\left(1/MIS_{i,t}\right)^{\lambda}}{\sum_{j=1}^{J}\left(1/MIS_{j.t}\right)^{\lambda}}$$

where $MIS_{i,t}$ 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 λ is close to zero, the combination reduces to the mean combination, whereas a large value for λ leads to the selection of the model with best historical accuracy. The parameter λ 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 $MIS_{i,t}$ 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 $MIS_{i,t}$ to be equal to the mean of $MIS_{i,t}$ 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 these 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.

	MIS					Mean rank				
Method	**All**	**U.S.**	**High**	**Med**	**Low**	**All**	**U.S.**	**High**	**Med**	**Low**
Mean	900	9799	1631	443	104	5.8	4	7.4	5.1	5.0
Median	778	11239	1172	435	110	5.8	9	6.4	5.0	5.9
Ensemble	794	11176	1211	448	112	6.7	8	7.1	6.4	6.6
Sym trim	878	10445	1515	446	109	6	6	6.3	5.8	5.8
Asym exterior trim	996	12447	1716	480	118	8.6*	11	9.4	7.9	8.5
Asym interior trim	895	9969	1625	429	97	4.8	5	7.0	4.1	3.3
Envelope	5284	64944	11037	1004	300	15*	16	14.8*	14.6*	15.5*
Previous best	877	11779	1217	637	135	9.8*	10	8.7	10.9*	9.9*
Inverse score	770	9130	1296	425	98	3.9	3	4.5	4.1	3.2
Inverse score tuning	677	8939	1000	441	104	5.2	2	4.6	5.8	5.4
Model 1	729	11174	948	512	114	6.3	7	5.3	7.9	5.7
Model 14	1918	30502	2936	909	227	13.7*	13	13.4*	13.5*	14.2*
Model 20	1957	33948	2808	933	248	13.5*	14	13.4*	13.6*	13.5*
Model 21	2004	35399	2846	949	254	14.3*	15	13.9*	14.5*	14.3*
Model 22	1014	19145	1312	535	130	9.2*	12	8.9	9.1	9.5*
Model 33	748	8681	1139	499	139	7.3*	1	5.1	7.5	9.8*
Mean MIS of models	2222	37870	3315	995	257					

* 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.

	MIS					Mean rank				
Method	**All**	**U.S.**	**High**	**Med**	**Low**	**All**	**U.S.**	**High**	**Med**	**Low**
Mean	3529	48497	6761	925	256	5.7	6	5.9	4.9	6.2
Median	3700	56185	6769	934	310	6.1	11	7.2	5.6	5.3
Ensemble	3588	50481	6747	941	318	6.5	8	7.1	6.1	6.2
Sym trim	3671	52563	6956	924	257	5.7	9	6.6	5.2	4.9
Asym exterior trim	3737	53839	7006	985	271	7.6*	10	8.1	7.2	7.3
Asym interior trim	3883	*36142*	8623	921	*208*	4.8	*1*	5.0	*4.7*	4.8
Envelope	6545	103456	11304	1968	663	11.7*	12	11.4*	11.7*	12.1*
Previous best	3531	49146	6371	1235	303	8.9*	7	7.7	10.2*	9.1*
Inverse score	3296	40479	6561	*920*	221	*4.2*	2	4.5	5.2	*3.1*
Inverse score tuning	*3100*	40758	*5918*	948	219	4.8	3	*4.2*	6.1	4.4
Model 21	7247	136492	11276	2202	658	12.2*	13	12.4*	12.1*	12.2*
Model 33	3457	44297	6552	1068	348	7.4*	4	5.3	8.0	9.2*
Mean MIS of models	6486	114911	10333	2114	633					

* 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.

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

* 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.

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

* 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.

	MIS					MAE				
Method	**All**	**U.S.**	**High**	**Med**	**Low**	**All**	**U.S.**	**High**	**Med**	**Low**
First 13-week period										
Mean	503	5992	978	166	42	56	1064	81	23	6
Median	419	6297	728	150	34	48	821	73	22	5
Model 1	450	7226	703	199	48	52	772	84	23	6
Second 13-week period										
Mean	452	7683	668	206	57	53	1004	68	27	9
Median	536	8361	791	278	79	55	1020	70	28	10
Ensemble	547	8570	803	286	79	55	1006	71	28	10
Sym trim	522	8756	745	267	71	54	1020	69	28	9
Asym exterior trim	586	11146	771	289	77	NA	NA	NA	NA	NA
Asym interior trim	468	7683	717	208	55	NA	NA	NA	NA	NA
Previous best	456	5649	638	313	112	53	826	74	30	10
Inverse score	416	7147	593	207	53	50	876	66	27	9
Inverse score tuning	416	7147	593	207	53	50	876	66	27	9
Model 1	373	6075	465	251	67	44	593	62	29	9
Model 33	463	5505	718	301	72	46	566	71	28	8
Third 13-week period										
Mean	1148	14161	2156	391	132	143	2874	189	60	20
Median	1168	18290	1956	405	135	143	2819	193	59	20
Ensemble	1203	17663	2066	434	142	143	2787	194	59	20
Sym trim	1216	15104	2290	399	141	144	2819	195	59	20
Asym exterior trim	1325	18193	2413	417	152	NA	NA	NA	NA	NA
Asym interior trim	1204	14325	2341	377	122	NA	NA	NA	NA	NA
Previous best	1307	21529	1903	671	158	167	3436	214	70	26
Inverse score	1029	12431	1914	377	125	139	2685	187	60	20
Inverse score tuning	838	11675	1343	396	139	128	2140	183	61	20
Model 1	1155	21363	1520	613	142	145	2857	193	64	18
Model 33	1051	14483	1634	530	198	121	1924	179	60	20
Fourth 13-week period										
Mean	1125	7176	2125	769	124	99	1181	183	41	10
Median	595	6391	682	646	116	56	953	72	34	9
Ensemble	594	6624	668	645	114	57	1001	71	34	9
Sym trim	890	6968	1499	701	114	85	953	160	35	9
Asym exterior trim	1078	7176	1987	765	124	NA	NA	NA	NA	NA
Asym interior trim	1022	7552	1828	739	116	NA	NA	NA	NA	NA
Previous best	856	7529	1076	965	133	74	1315	96	42	10
Inverse score	875	7551	1389	726	118	96	1131	178	40	10
Inverse score tuning	782	7768	1055	758	121	66	994	92	40	10
Model 1	638	5108	828	690	133	55	829	75	36	10
Model 33	722	5782	1036	686	146	70	1363	87	37	11

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.

	MIS					MAE				
Method	**All**	**U.S.**	**High**	**Med**	**Low**	**All**	**U.S.**	**High**	**Med**	**Low**
First 13-week period										
Mean	1568	14137	3516	301	148	140	1863	258	46	15
Median	1811	14288	4268	275	156	144	2318	250	42	13
Model 1	1686	12850	3992	267	141	140	2034	252	43	14
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.

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

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.

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

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.

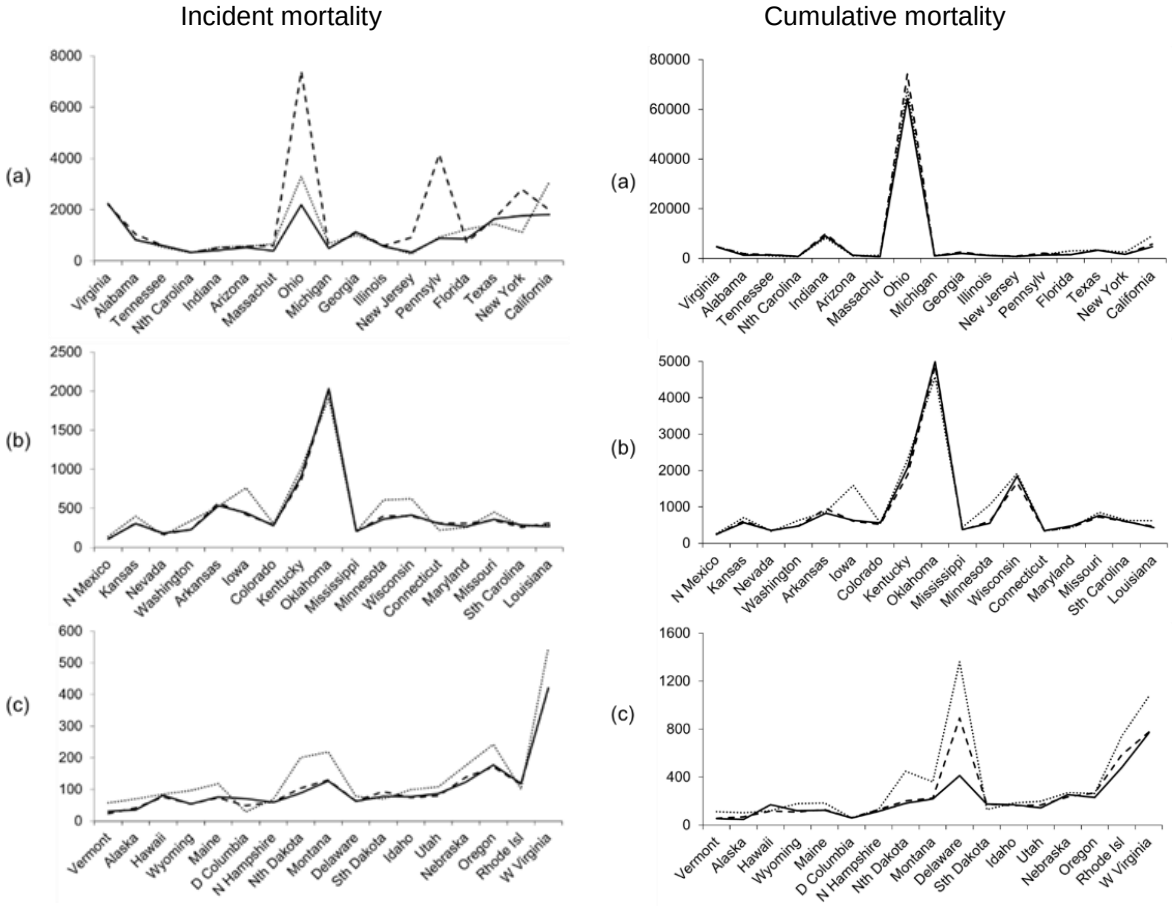

Fig. 5 MIS for 95% interval forecasts of incident and cumulative mortality at (a) high, (b) medium and (c) low mortality U.S. locations for the inverse score with tuning combining (black line), mean combining (dashed line) and Model 33 (dotted line).

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).

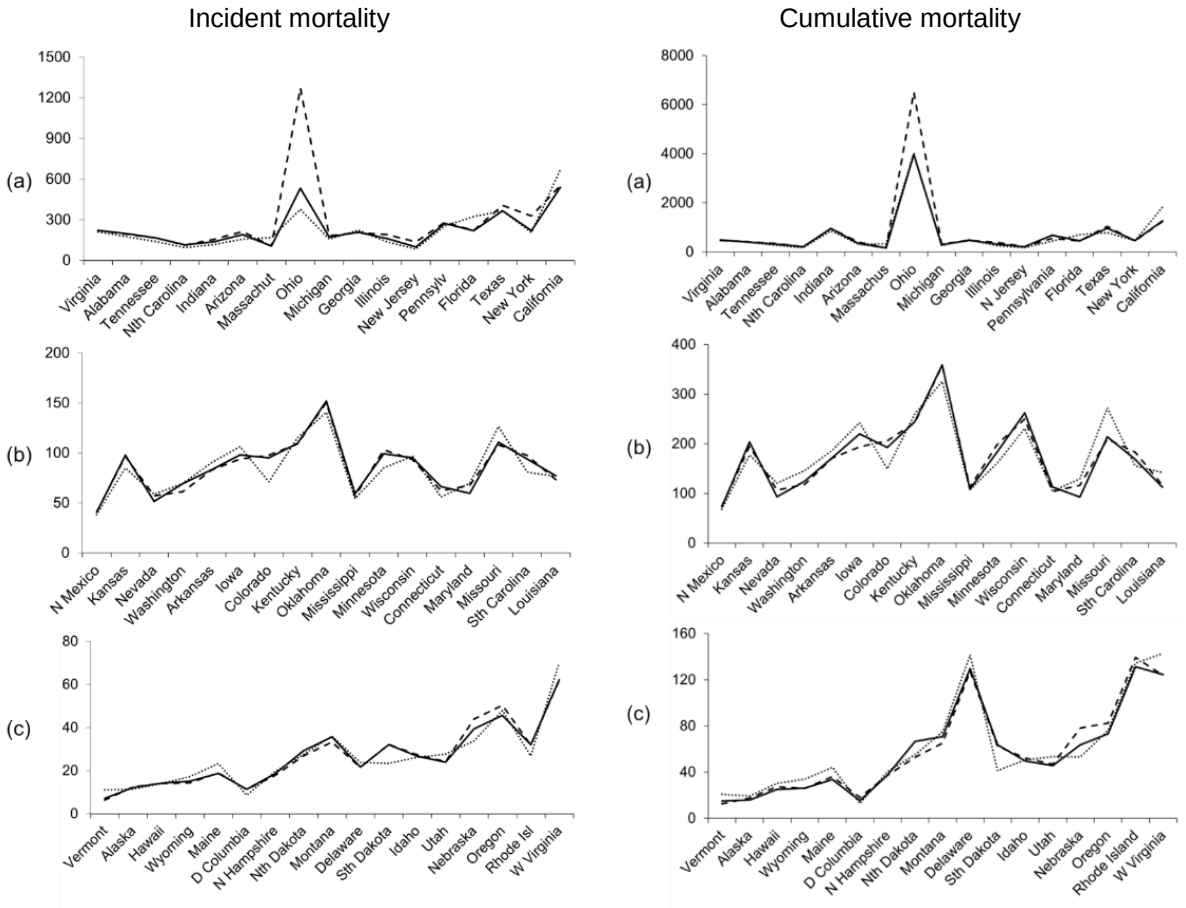

Fig. 6 MAE for point forecasts of incident and cumulative mortality at (a) high, (b) medium and (c) low mortality U.S. locations, for the inverse score with tuning combining (black line), mean combining (dashed line) and Model 33 (dotted line).

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%.

	Incident mortality					Cumulative mortality				
Method	**All**	**U.S.**	**High**	**Med**	**Low**	**All**	**U.S.**	**High**	**Med**	**Low**
Mean	87.2	93.4	87.2	89.7	84.4	85.7	79.6	81.5	90.4	85.5
Median	85.6	89.3	85.6	88.1	83.0	82.4	71.6	78.8	87.9	81.0
Ensemble	85.3	88.6	84.9	87.7	83.0	82.2	71.6	79.0	87.4	80.8
Sym trim	86.6	89.3	87.0	88.8	83.8	84.2	74.9	80.1	88.9	84.0
Asym exterior trim	81.9	86.6	82.0	84.5	79.0	80.9	72.2	76.9	85.5	80.8
Asym interior trim	91.9	94.0	91.4	93.7	90.3	91.6	85.5	90.5	92.9	91.9
Envelope	98.7	100.0	99.6	99.5	96.9	98.1	99.4	96.9	99.3	97.9
Previous best	83.7	89.3	84.9	84.9	81.1	80.3	84.2	79.8	81.6	79.3
Inverse score	90.0	98.7	90.1	92.8	86.7	88.6	82.3	84.8	92.6	88.7
Inverse score tuning	89.7	98.7	90.5	91.7	86.4	88.0	86.9	85.2	91.3	87.5
Model 1	91.7	98.0	93.6	93.1	88.1	NA	NA	NA	NA	NA
Model 14	66.4	85.4	63.8	69.7	64.5	NA	NA	NA	NA	NA
Model 20	67.2	82.9	63.7	67.8	69.3	NA	NA	NA	NA	NA
Model 21	67.2	82.9	64.0	68.0	68.8	61.8	69.7	56.6	62.2	66.3
Model 22	87.4	77.4	87.1	88.6	87.1	NA	NA	NA	NA	NA
Model 33	83.9	84.8	87.2	85.7	78.9	81.8	72.4	82.2	85.6	78.1

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 week 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 have 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

1. Phelan AL, Katz R, Gostin LO. The Novel Coronavirus Originating in Wuhan, China: Challenges for Global Health Governance. Jama. 2020;323(8):709-10.
2. Jewell NP, Lewnard JA, Jewell BL. Predictive Mathematical Models of the COVID-19 Pandemic: Underlying Principles and Value of Projections. Jama. 2020;323(19):1893-4.
3. Looi M-K. Covid-19: Is a second wave hitting Europe? BMJ. 2020;371:m4113.
4. Melnick ER, Ioannidis JPA. Should governments continue lockdown to slow the spread of covid-19? BMJ. 2020;369:m1924.
5. Policy brief: Education during COVID-19 and beyond [press release]. 2020.

6. Wise J. Covid-19: Experts divide into two camps of action—shielding versus blanket policies. BMJ. 2020;370:m3702.
7. Adam D. Special report: The simulations driving the world's response to COVID-19. Nature. 2020;580(7803):316-8.
8. Holmdahl I, Buckee C. Wrong but Useful — What Covid-19 Epidemiologic Models Can and Cannot Tell Us. New England Journal of Medicine. 2020;383(4):303-5.
9. Ioannidis JPA, Cripps S, Tanner MA. Forecasting for COVID-19 has failed. International Journal of Forecasting. 2020.
10. Buckee CO, Johansson MA. Individual model forecasts can be misleading, but together they are useful. Eur J Epidemiol. 2020;35(8):731-2.
11. George EPB. Science and Statistics. Journal of the American Statistical Association. 1976;71(356):791-9.
12. Ferguson N, Laydon D, Nedjati-Gilani G, al e. Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand28 Dec 2020. Available from: http://hdl.handle.net/10044/1/77482.
13. Panovska-Griffiths J. Coronavirus: we've had 'Imperial', 'Oxford' and many more models – but none can have all the answers2020 28 Dec 2020. Available from: https://theconversation.com/coronavirus-weve-had-imperial-oxford-and-many-more-models-but-none-can-have-all-the-answers-135137.
14. Surowiecki J. The Wisdom of Crowds: Why the Many are Smarter Than the Few and how Collective Wisdom Shapes Politics,Business, Economies, Societies, and Nations: Doubleday & Co; 2004.
15. Claeskens G, Magnus JR, Vasnev AL, Wang W. The forecast combination puzzle: A simple theoretical explanation. International Journal of Forecasting. 2016;32(3):754-62.
16. Smith J, Wallis KF. A Simple Explanation of the Forecast Combination Puzzle*. Oxford Bulletin of Economics and Statistics. 2009;71(3):331-55.
17. Bates JM, Granger CWJ. The Combination of Forecasts. OR. 1969;20(4):451-68.
18. Busetti F. Quantile Aggregation of Density Forecasts. Oxford Bulletin of Economics and Statistics. 2017;79(4):495-512.
19. Timmermann A. Chapter 4 Forecast Combinations. In: Elliott G, Granger CWJ, Timmermann A, editors. Handbook of Economic Forecasting. 1: Elsevier; 2006. p. 135-96.
20. Krishnamurti TN, Kishtawal CM, LaRow TE, Bachiochi DR, Zhang Z, Williford CE, et al. Improved Weather and Seasonal Climate Forecasts from Multimodel Superensemble. Science. 1999;285(5433):1548-50.
21. Leutbecher M, Palmer TN. Ensemble forecasting. J Comput Phys. 2008;227(7):3515–39.
22. Johansson MA, Apfeldorf KM, Dobson S, Devita J, Buczak AL, Baugher B, et al. An open challenge to advance probabilistic forecasting for dengue epidemics. Proceedings of the National Academy of Sciences. 2019;116(48):24268-74.
23. Ray EL, Wattanachit N, Niemi J, Kanji AH, House K, Cramer EY, et al. Ensemble Forecasts of Coronavirus Disease 2019 (COVID-19) in the U.S. medRxiv. 2020:2020.08.19.20177493.
24. Reich NG, McGowan CJ, Yamana TK, Tushar A, Ray EL, Osthus D, et al. Accuracy of real-time multi-model ensemble forecasts for seasonal influenza in the U.S. PLoS Comput Biol. 2019;15(11):e1007486-e.
25. Yamana TK, Kandula S, Shaman J. Individual versus superensemble forecasts of seasonal influenza outbreaks in the United States. PLoS Comput Biol. 2017;13(11):e1005801.
26. Taylor JW, Taylor KS. Combining Probabilistic Forecasts of COVID-19 Mortality in the United States European Journal of Operational Research. European Journal of Operational Research. 2021;[in press].
27. Bracher J, Ray E, Gneiting T, Reich NG. Evaluating epidemic forecasts in an interval format. arXiv: Applications. 2020.
28. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis. 2020;20(5):533-4.
29. Gneiting T, Raftery AE. Strictly Proper Scoring Rules, Prediction, and Estimation. Journal of the American Statistical Association. 2007;102(477):359-78.
30. Winkler RL, Grushka-Cockayne Y, Jr. KCL, Jose VRR. Probability Forecasts and Their Combination: A Research Perspective. Decision Analysis. 2019;16(4):239-60.
31. Diebold FX, Mariano RS. Comparing Predictive Accuracy. Journal of Business & Economic Statistics. 2002;20(1):134-44.

32. Koning AJ, Franses PH, Hibon M, Stekler HO. The M3 competition: Statistical tests of the results. International Journal of Forecasting. 2005;21(3):397-409.
33. Gaba A, Tsetlin I, Winkler RL. Combining Interval Forecasts. Decision Analysis. 2017;14(1):1-20.
34. Hora SC, Fransen BR, Hawkins N, Susel I. Median Aggregation of Distribution Functions. Decision Analysis. 2013;10(4):279-91.
35. Jose VRR, Grushka-Cockayne Y, Lichtendahl KC. Trimmed Opinion Pools and the Crowd's Calibration Problem. Management Science. 2014;60(2):463-75.
36. Park S, Budescu D. Aggregating multiple probability intervals to improve calibration. Judgment and decision making. 2015;10:130-43.
37. Capistrán C, Timmermann A. Forecast Combination With Entry and Exit of Experts. Journal of Business & Economic Statistics. 2009;27(4):428-40.
38. Borenstein M, Hedges L, Higgins J, Rothstein H. Introduction to Meta-Analysis: John Wiley & Sons.
39. Grushka-Cockayne Y, Jose VRR. Combining prediction intervals in the M4 competition. International Journal of Forecasting. 2020;36(1):178-85.
40. McGowan CJ, Biggerstaff M, Johansson M, Apfeldorf KM, Ben-Nun M, Brooks L, et al. Collaborative efforts to forecast seasonal influenza in the United States, 2015–2016. Scientific Reports. 2019;9(1):683.
41. Yeh RW. Academic Cardiology and Social Media: Navigating the Wisdom and Madness of the Crowd. Circ Cardiovasc Qual Outcomes. 2018;11(4):e004736.
42. Sridhar D, Majumder MS. Modelling the pandemic. BMJ. 2020;369:m1567.
43. Lipsitch M, Santillana M. Enhancing Situational Awareness to Prevent Infectious Disease Outbreaks from Becoming Catastrophic. Curr Top Microbiol Immunol. 2019;424:59-74.

SUPPLEMENTARY INFORMATION

Fig S1 Numbers of weekly incident COVID-19 deaths in U.S. locations between weeks ending 23 May 2020 and 15 May 2021

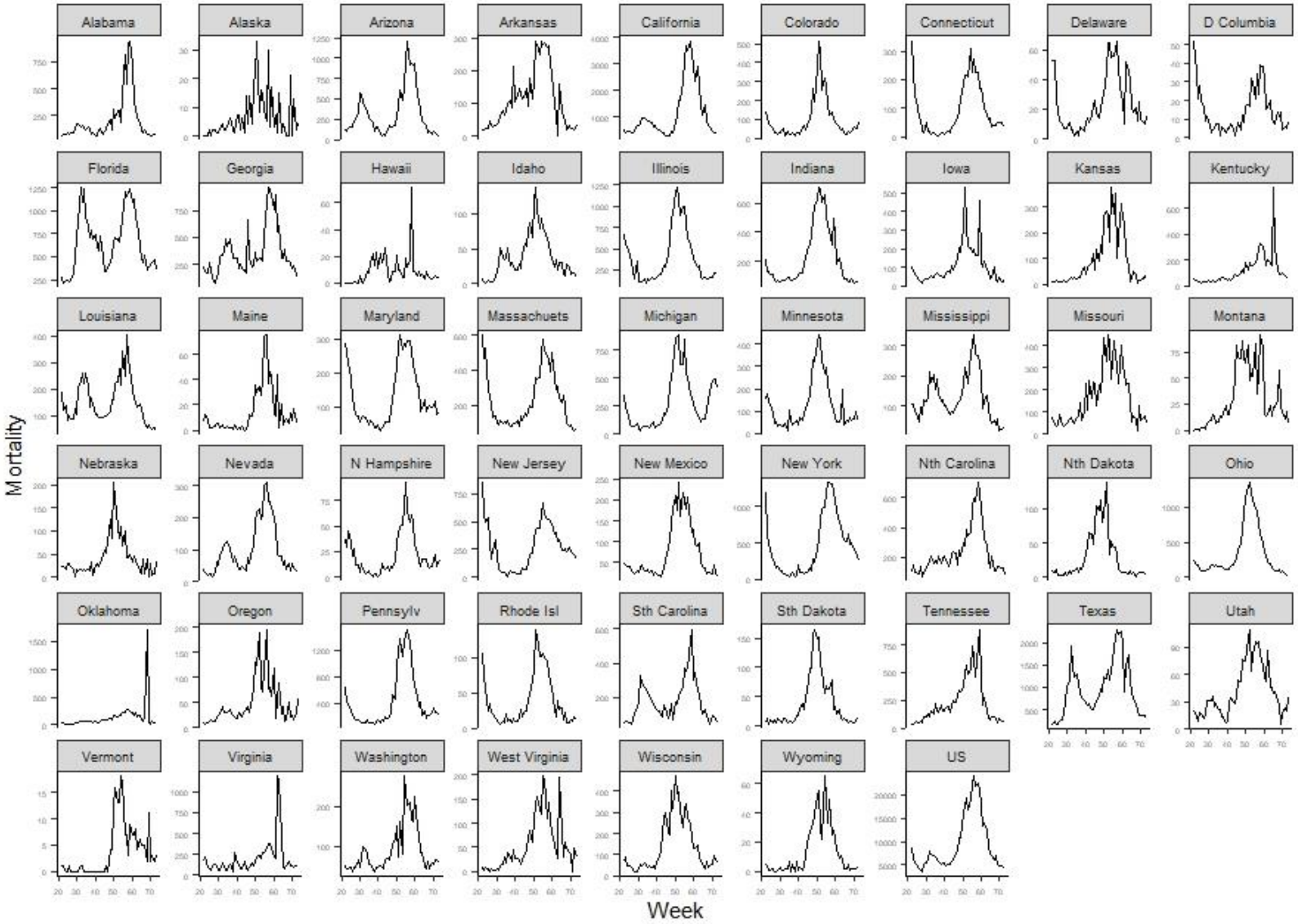

Fig S2 Numbers of weekly cumulative COVID-19 deaths in U.S. locations between weeks ending 23 May 2020 and 15 May 2021

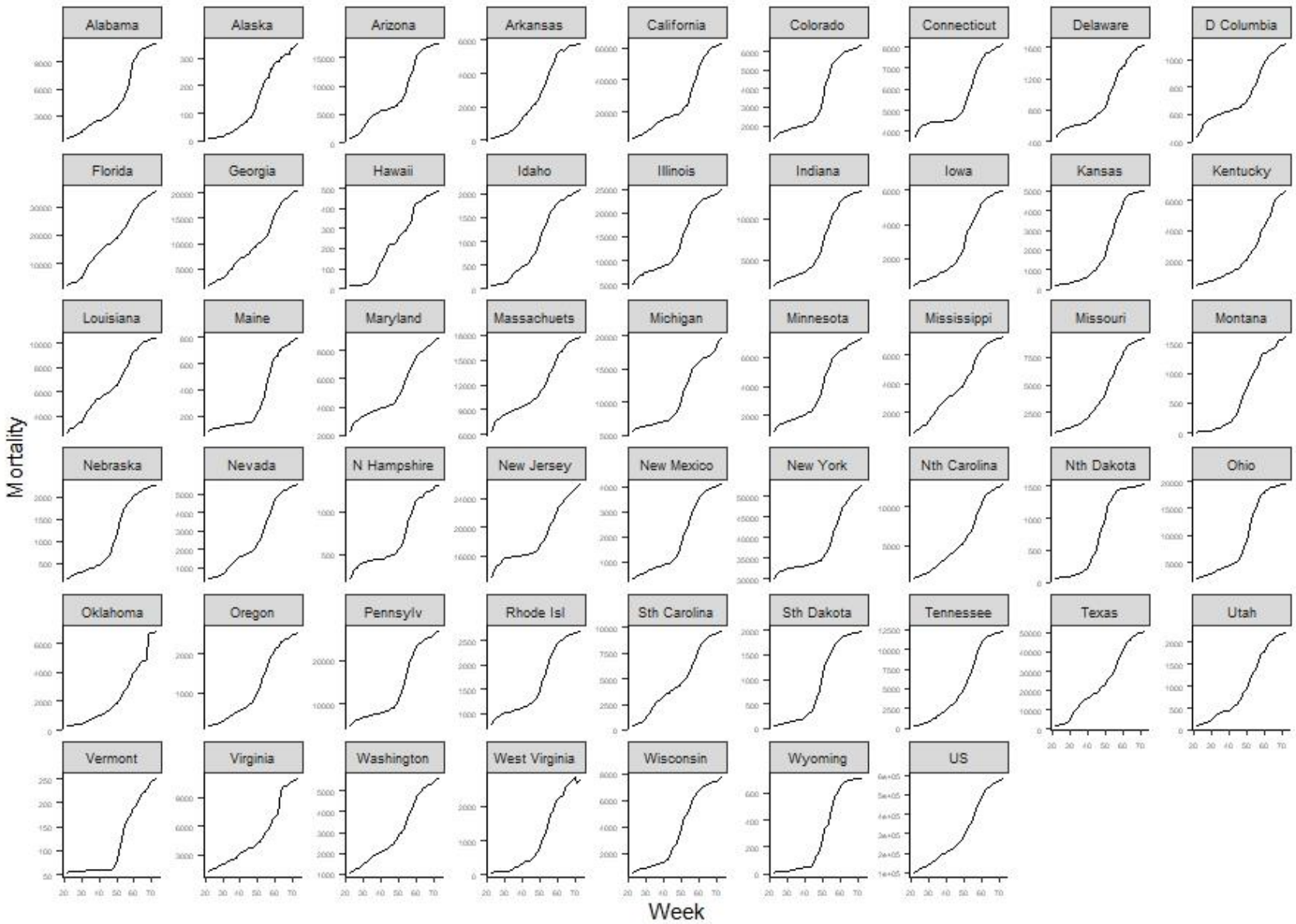

Table S1 Frequency of being one of the top three methods in the columns for the scores in Tables 1 to 4

		Mortality		Forecast type	
Method	**Total**	**Incident**	**Cumulative**	**Interval**	**Point**
Mean	2	2	0	1	1
Median	6	4	2	1	5
Ensemble	6	3	3	0	6
Symm trim	6	2	4	1	5
Asym exterior trim	0	0	0	0	NA
Asym interior trim	5	2	3	5	NA
Envelope	0	0	0	0	NA
Previous best	2	0	2	1	1
Inverse score	9	4	5	7	2
Inverse score tuning	16	7	9	8	8
Model 1	6	6	NA	2	4
Model 14	0	0	NA	0	0
Model 20	0	0	NA	0	0
Model 21	0	0	0	0	0
Model 22	0	0	NA	0	0
Model 33	11	7	4	5	6

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

		Mortality		Forecast type	
Method	**Total**	**Incident**	**Cumulative**	**Interval**	**Point**
Mean	2	1	1	2	0
Median	7	5	2	1	6
Ensemble	5	3	2	0	5
Symm trim	5	1	4	1	4
Asym exterior trim	0	0	0	0	NA
Asym interior trim	8	3	5	8	NA
Envelope	0	0	0	0	NA
Previous best	1	0	1	0	1
Inverse score	12	6	6	10	2
Inverse score tuning	13	5	8	7	6
Model 1	0	0	NA	0	0
Model 14	3	3	NA	0	3
Model 20	0	0	NA	0	0
Model 21	0	0	0	0	0
Model 22	0	0	NA	0	0
Model 33	7	5	2	2	5

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.

Method	MIS					MAE				
	All	**U.S.**	**High**	**Med**	**Low**	**All**	**U.S.**	**High**	**Med**	**Low**
1 week ahead										
Mean	687	6877	1224	386	88	68	1202	92	34	11
Median	618	8679	914	370	95	64	1152	86	33	11
Ensemble	621	8541	927	374	95	65	1181	85	33	11
Sym trim	638	7222	1075	364*	88†	65*	1152	87	32*	11
Asym interior trim	647	6999	1125	361	81*†	68	1223	91	33	11
Inverse score	594*	6506	988*	364*	84*†	66*	1159	89	33	11
Inverse score tuning	533*	6245	807	370	87†	62*	1032	85*	34*	11*
Model 21	527*	6005	770†	394	94	62	1039	84	33	10
Model 33	710	8352	1121	449	112	66	1081	93	34	10
2 weeks ahead										
Mean	892	9345	1644	436	97	99	1486	162	40	13
Median	705	9795	1072	407	100	76	1385	102	38*	12*
Ensemble	713	9739	1087	418	101	77	1426	102	38*	12
Sym trim	841	9761	1477	424	98	94*	1390	156	38*	12*
Asym interior trim	913	9322	1743	412	91	99	1514	161	39	13
Inverse score	742*	8069	1285*	418	91*	96	1385*	161	39	12
Inverse score tuning	635*	7693	958	434	96	75	1162*	108	40	12
Model 21	645	8660	888	473	102	74	1237	103	40	12
Model 33	720	9218	1088	454	120	***73***	1155	105	38	12
3 weeks ahead										
Mean	979	9660	1863	459	105	107	1809	163	46	14
Median	799	12427	1160	446	107	92	1726	124	43*	14
Ensemble	813	11915	1217	459	108	92	1708	124	44	14
Sym trim	953	11156	1694	456	110	103	1722	157	44	14
Asym interior trim	942	9921	1751	447	100	107	1825	162	45	14
Inverse score	815*	8952	1432*	436	100*	103*	1666*	159	45	14*
Inverse score tuning	686	8912	1018	452	105	88	1451	124	46	14
Model 21	806	13889	988	548	113	90	1632	119	46	13
Model 33	698	6844	1097	495	140	80	1191	119	43	14
4 weeks ahead										
Mean	1044	13315	1792	491	126	121	2347	165	52	16
Median	989	14055	1543	518	138	112	2263	143	49	15
Ensemble	1029	14509	1611	539	143	111	2204	145	50	15
Sym trim	1077	13641	1814	542	138	118	2263	163	51	15
Asym interior trim	1076	13634	1879	495	116*	121	2353	165	52	16
Inverse score	929*	12993	1478*	481	117	116*	2130*	162	51	16
Inverse score tuning	854†	12906	1216	508	128	104*	1777	145	53	15
Model 21	939	16141	1145	632	146	106	1931	141	55	14
Model 33	863	10312	1249	598	185	99	1680	137	51	17

* 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.

	MIS					MAE				
Method	**All**	**U.S.**	**High**	**Med**	**Low**	**All**	**U.S.**	**High**	**Med**	**Low**
1 week ahead										
Mean	3319	62108	5666	585	249	160	3541	223	40	17
Median	3123	57217	5473	451*	263	144*	3036	209*	36*	16*
Ensemble	3077	54979	5461	453*	265	143*†	2979	209*	36*	16*
Sym trim	3277	63747	5560	483*	231†	148	3190	213	37*	16*
Asym interior trim	2204*†	33641*†	4136*†	491*	136*†	156	3402	220	39	17*
Previous best	2559*†	31436†	5282	532	165*†	147	2767*	228	44	17
Inverse score	3195	55789*	5681	612	199*†	159	3548	222	40	16*
Inverse score tuning	2736*†	39628†	5290*	574	175*†	158	3520	219	40	16*
Model 33	3271	62685	5503	540	276	142*	2858	213	37	16
2 weeks ahead										
Mean	3045	38620	6102	720	219†	223	4031	357	64	24
Median	3303	47609	6284	737	282	200	3879	297	64	24
Ensemble	3231	44870	6220	740	285	224	4037	359	64	24
Sym trim	3105	39643	6213	728	223†	202	3983	297	63	24
Asym interior trim	3491	32197	7897	729	158*†	224	4037	359	64	24
Previous best	3063	40956	5797	928	235	204	3354	321	81	26
Inverse score	2842*†	31725	5927*	713	188*†	217	3898*	347	64	24*
Inverse score tuning	2756†	32620	5601†	738	171†	193	3651	289	64	24
Model 33	3332	48704	6166	856	305	197	3601	300	67	24
3 weeks ahead										
Mean	3594	43109	7178	1022	257†	303	5022	497	100	33
Median	3795	52918	7100	1077	318	265	4769	400	98	34
Ensemble	3655	46027	7070	1075	329	263	4674	397	98	34
Sym trim	3777	48037	7423	1041	263†	270	4996	399	98	33
Asym interior trim	4503	36011	10377	1050	230†	305	5058	500	101	34
Previous best	3766	53485	6604	1425	344	277	4129	442	127	37
Inverse score	3349*	35893	6899*	1012	221*†	292	4763*	480	100	33*
Inverse score tuning	3177	39846	6090†	1052	232†	255	4378	389	99	33
Model 33	3442	35365	6859	1229	361	257	4220	403	102	*34*
4 weeks ahead										
Mean	4158	50151	8096	1373	300†	395	6422	642	143	44
Median	4580	66996	8219	1471	379	350	6070	530	140	44
Ensemble	4387	56047	8235	1495	393	348	6016	526	140	45
Sym trim	4526	58825	8629	1444	312†	355	6387	527	140	44†
Asym interior trim	5335	42720	12085	1413	308	398	6503	647	144	44
Previous best	4734	70708	7799	2053	470	381	5654	600	183	50
Inverse score	3799*	38510	7736*	1345	275*†	381	6095*	619	143	43*
Inverse score tuning	3731	50936	6690†	1428	298†	337	5617	514	142	44
Model 33	3781	30433	7680	1646	450	328	4963	522	145	46

* 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.

	MIS					Mean rank				
Method	**All**	**U.S.**	**High**	**Med**	**Low**	**All**	**U.S.**	**High**	**Med**	**Low**
Mean	789	9799	1165	324	69	5.7	4	6.3	4.7	4.3
Median	723	11239	895	327	73	5.8	9	5.5	4.8	4.9
Ensemble	737	11176	925	339	74	6.7	8	6.0	6.3	5.3
Sym trim	758	10445	1038	327	72	5.8	6	5.1	5.4	4.8
Asym exterior trim	888	12447	1231	359	79	8.4	11	7.9	7.2	7.1
Asym interior trim	816	9969	1254	304	63	4.7	5	6.1	3.3	3.0
Envelope	5042	64944	8718	885	220	15.1	16	12.9	14.1	12.8
Previous best	864	11779	1066	493	87	9.8	10	7.9	10.1	7.8
Inverse score	686	9130	950	305	64	3.7	3	3.7	3.4	2.8
Intervals score tuning	651	8939	846	322	68	5.2	2	4.4	5.2	4.5
Model 22	993	19145	1054	421	85	9.3	12	7.9	8.8	7.7
Model 21	724	11174	840	394	67	6.3	7	5.1	7.6	3.9
Model 14	1961	30502	2546	812	153	14.1	13	12.0	13.5	11.9
Model 20	2005	33948	2496	768	164	13.6	14	12.2	12.8	11.0
Model 21	2056	35399	2531	781	170	14.3	15	12.6	13.6	11.7
Model 33	705	8681	913	386	97	7.5	1	4.5	7.3	8.5
Mean MIS of models	2208	37870	2705	856	186					

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.

	MIS					Mean rank				
Method	**All**	**U.S.**	**High**	**Med**	**Low**	**All**	**U.S.**	**High**	**Med**	**Low**
Mean	2003	48497	1806	637	124	5.6	6	5.1	4.4	5.3
Median	2257	56185	2001	679	124	5.8	11	6.1	5.4	3.6
Ensemble	2145	50481	2021	686	126	6.3	8	6.1	6.0	4.4
Sym trim	2141	52563	1930	651	121	5.4	9	5.5	4.9	3.8
Asym exterior trim	2219	53839	2013	687	138	7.6	10	7.1	6.5	6.3
Asym interior trim	1741	36142	1831	634	120	4.8	1	4.2	4.2	4.6
Envelope	4738	103456	4480	1750	506	12.4	12	10.5	11.6	10.6
Previous best	2252	49146	2095	937	170	9.1	7	7.1	9.5	7.7
Inverse score	1778	40479	1681	635	114	4.2	2	3.8	4.8	2.6
Inverse score tuning	1765	40758	1603	654	121	5.0	3	3.9	5.5	3.9
Model 21	6015	136492	6096	1793	358	12.2	13	11.1	11.3	9.9
Model 33	2074	44297	2047	798	161	7.5	4	4.9	7.8	7.4
Mean MIS of models	5229	114911	5216	1789	384					

Appendix Forecasting models

Contributors	**Short model name**	**Model description***	**Access and licencing information Citations**
Wattanachit N, Ray EL, Reich N	COVID hub-ensemble	An ensemble, or model average, of submitted forecasts to the COVID-19 Forecast Hub.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/COVIDhub-ensemble https://www.medrxiv.org/content/10.1101/2020.08.19.20177493v1
COMPARTMENTAL			
Tomar V, Jain C	Auquan-SEIR†	Modified SEIR model with compartments for reported and unreported infections. Non-linear mixed effects curve-fitting.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/Auquan-SEIR
Panano B.	BPangano-RtDriven	Projects infections and deaths for 223 locations using an SIR model.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/BPagano-RtDriven https://bobpagano.com/covid-19-modeling/
Carlson E, Henderson M, Kelly C, Kofman I, Zhang X	CovidActNow-SEIR_CAN	SEIR model forecasts of cumulative deaths, incident deaths, incident hospitalizations by fitting predicted cases, deaths, and hospitalizations to the observations.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/CovidActNow-SEIR_CAN
Li ML, Bouardi HT, Lami OS, Trikalinos TA, Trichakis NK, Bertsimas D	CovidAnalytics-DELPHI	SEIR model augmented with underdetection and interventions. Projections account for reopening and assume interventions would be re-enacted if cases continue to climb.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/CovidAnalytics-DELPHI https://www.covidanalytics.io/DELPHI_documentation_pdf
Chhatwal J, Ayer T, Linas B, Dalgic O, Mueller P, Adee M, Ladd MA, Xiao J	Covid19Sim-Simulator	An interactive tool that uses a validated SEIR compartment model.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/Covid19Sim-Simulator
Pei S, Yamana T, Kandula S, Yang W, Galanti M, Shaman J	CU-select	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.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/CU-select https://doi.org/10.1101/2020.03.21.20040303 https://www.medrxiv.org/content/10.1101/2020.05.04.20090670v2
Pei S, Yamana T, Kandula S, Yang W, Galanti M, Shaman J	CU-nochange	This metapopulation county-level SEIR model assumes that current contact rates will	https://github.com/reichlab/covid19-forecast-

		remain unchanged in the future.	hub/tree/master/data-processed/CU-nochange https://doi.org/10.1101/2020.03.21.20040303
Max A, Epshteyn A, Kang B, Li C-L, Sava D, Parish D, Miller D, Kanal E, Liu H, Nakhost H, Jones I, Lai J, Repenning J, Yoon J, Ramasamy K, Zhang L, Le L, Nikoltchev M, Siegler M, Dusenberry M, Yoder N, Rozenfeld O, Rangaswamy P, Sinha R, Xie R, Arik S, Singh S, Tsai T, Pfister T, Menon V, Karande V, Y, Li Y	Google-Harvard-CPF	Our model improves upon standard compartmental models by using temporally and spatially rich data, and integrating covariate encodings into compartment transitions via end-to-end learning.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/Google_Harvard-CPF https://arxiv.org/abs/2008.00646
Lemaitre JC, Bi Q , Hulse JD, Grabowski MK, Grantz KH, Kaminsky J, Lauer SA, Lee EC, Meredith HR, Perez-Saez J, Truelove SA, Keegan LT, Kaminsky K, Shah S, Wills J, Aquilanti P-Y, Raman K, Subramaniyan A, Thursam G, Tran A.	JHU_IDD-CovidSP	County-level metapopulation model with commuting and stochastic SEIR disease dynamics with social-distancing indicators.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/JHU_IDD-CovidSP https://doi.org/10.1038/s41598-021-86811-0
Kinsey M, Tallaksen K, Obrecht RF, Asher L, Costello C, Kelbaugh M, Wilson S	JHUAPL_Bucky	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.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/JHUAPL-Bucky
Baek J, Farias V, Georgescu A, Levi R, Sinha D, Wilde J, Zheng A	MITCovAlliance-SIR	SIR model trained on public heath regions. SIR parameters are functions of static demographic and time-varying mobility features. Two-stage approach that first learns magnitude of peak infections.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/MITCovAlliance-SIR https://arxiv.org/abs/2006.06373
Vespignani A, Chinazzi M, Davis JT, Mu K, Pastore y Piontti A, Samay N, Xiong X, Halloran ME, Longini IM,	MOBS-GLEAM_COVID	Metapopulation, age structured SLIR model. Superimposed on the worldwide population and mobility layers is an agent-based epidemic model that	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/MOBS-GLEAM_COVID

Dean NE, Viboud C, Sun K, Litvinova M, Gioannini C, Rossi L, Ajelli M		defines the infection and population dynamics. Makes predictions about the future that are dependent on the assumption that current interventions continue.	https://uploads-ssl.webflow.com/58e6558acc00ee8e4536c1f5/5e8bab44f5baae4c1c2a75d2_GLEAM_web.pdf
Gao Z, Li C, Zheng S, Bian J, Xie X, LiuT-Y	MSRA-DeepST	A deep spatio-temporal network with knowledge based SEIR as a regularier under the assumption of spatio-temporal process in pandemic of different regions.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/MSRA-DeepST
Espana G, Oidtman R, Cavany S, Costello A, Wieler A, Lerch A, Barbera C, Poterek M, Tran Q, Moore S, Perkins A	NotreDame-Mobility	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.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/NotreDame-mobility
Koyluoglu U, Milliken J	OliverWyman-Navigator	Forecasts and scenario analysis for Detected and Undetected cases and death counts following a compartmental formulation with non-stationary transition rates.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/OliverWyman-Navigator
Turtle J, Ben-Nun M, Riley P	PSI-DRAFT	A stochastic/deterministic, single-population SEIRX model that stratifies by both age distribution and disease severity and includes generic intervention fitting.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/PSI-DRAFT
Shi Y, Shah T, Ban X	RPI-UW-Mob_Collision	A mobility-informed simplified SIR model motivated by collision theory.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/RPI-UW-Mob-Collision https://www.medrxiv.org/content/10.1101/2020.07.25.20162016v1
Snyder TL, Wilson DD	SWC-TerminusCM	Mechanistic compartmental model using disease parameter estimates from literature. It uses Bayesian inference to predict the most likely model parameters.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/SWC-TerminusCM
Cobey S, Arevalo P, Baskerville E, Carran S, Gostic K, McGough L, Ranjeva S, Wen F	UChicago-COVIDIL	Compartmental, age-structured SEIR model that infers past SARS-CoV-2 transmission rates and forecasts mortality under current and hypothetical public health interventions.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/UChicago-CovidIL

Gu Q, Xu P, Chen J, Wang L, Zou D, Zhang W	UCLA-SuEIR	Variant of the SEIR model considering both untested and unreported cases. The model considers reopening and assumes susceptible population will increase after the reopen.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/UCLA-SuEIR https://www.medrxiv.org/content/10.1101/2020.05.24.20111989v1
Chen YQ, Zhao Y, Guo L	UCM-MESALab-FoGSEIR	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.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/UCM_MESALab-FoGSEIR
Sheldon D, Gibson G, Reich N	UMass-MechBayes	Bayesian compartmental model with observations on cumulative case counts and cumulative deaths. Model is fit independently to each state. Model includes observation noise and a case detection rate.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/UMass-MechBayes
Mayo ML, Rowland MA, Parno MD, Detwiller ID, Farthing MW, England WP George GE	USACE-ERDC_SEIR	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.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/USACE-ERDC_SEIR
Jain S, Tiwari A, Deva A, Kulkarni M, Shingi S, Bannur N, White J, Merugu S, Raval A	Wadhwani_AI-BayesOpt	A novel model-agnostic Bayesian optimization ("BayesOpt") approach for learning the parameters of our SEIR model from observed data.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/Wadhwani_AI-BayesOpt
Gu Y	YYG-ParamSearch	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.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/YYG-ParamSearch https://covid19-projections.com/about/
NON-COMPARTMENTAL			
O'Dea E	CEID-Walk	A random walk model with drift. A least squares line is fitted to the tail observations of a target time series to estimate the drift and step variance of a random walk model.	https://github.com/reichlab/covid19-forecast-hub/blob/master/data-processed/CEID-Walk/metadata-CEID-Walk.txt

Green A, Hu A, Jahja M, Ventura V, Wasserman L, Tibshirani Rob, Shankar V, Bien J, Brooks L, Narasimhan B, Rajanala S, Rumack A, Simon N, Sharpnack J, McDonald D(University of British Columbia), Ryan Tibshirani (Senior author, and the Delphi COVID-19 Response Team	CMU-Timeseries §	A basic AR-type time series model fit using case counts and deaths as features.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/CMU-TimeSeries
Wang Y, Zeng D, Wang Q, Xie S	Columbia_UNC-SurvCon	Survival-convolution model with piece-wise transmission rates that incorporates latent incubation period and provides time-varying effective reproductive number.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/Columbia_UNC-SurvCon https://www.frontiersin.org/article/10.3389/fpubh.2020.00325
Ray EL, Tibshirani R	COVIDhub-baseline	Baseline prediction model.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/COVIDhub-baseline
Kalantari R, Zhou M.	DDS-NBDS	Jointly modeling daily deaths and cases using a negative binomial distribution based nonparametric Bayesian generalized linear dynamical system.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/DDS-NBDS https://dds-covid19.github.io/
Sherratt K, Bosse N, Abbott S, Hellewell J, Meakin S, Munday J, Funk S	epiforecasts-ensemble1	A deaths forecast using the renewal equation and time-series forecasts of the time-varying reproduction number.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/epiforecasts-ensemble1 https://doi.org/10.12688/wellcomeopenres.16006.1
Keskinocak P, Aglar BEO, Baxter A, Asplund J, Serban N	GT_CHHS-COVID19	Agent-based simulation model to project COVID19 infection spread.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/GT_CHHS-COVID19
Prakash BA, Rodriguez A, Cui J, Tabassum A, Adhikari B, Sun J, Xiao D, Qiang C	GT-DeepCOVID	Data-driven approach based on deep learning for forecasting mortality and hospitalizations using syndromic, clinical, demographic, mobility and point-of-care data.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/GT-DeepCOVID

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 https://www.medrxiv.org/content/10.1101/2020.03.27.20043752v1
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 https://arxiv.org/abs/2004.14103
Chiang W-H, Mohler G	IUPUI-HkPrMobiDyR	Hawkes processes with Dynamic reproduce number.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/IUPUI-HkPrMobiDyR https://doi.org/10.1101/2020.06.06.20124149
Marshall M, Gardner L, Drew C, Burman E, Nixon K	JHU_CSSE-DECOM	County-level, empirical machine learning model driven by epidemiological, mobility, demographic, and behavioral data.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/JHU_CSSE-DECOM
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 https://arxiv.org/abs/2007.07156
Osthus D, Del Valle S, Manore C, Weaver B, Castro L, Shelley S, Smith M, Spencer J, Fairchild G, Travis Pitts T, Gerts D, Dauelsberg L, Daughton A, Gorris M, Hornbein B, Israel D, Parikh N, Shutt D, Ziemann A	LANL-GrowthRate	Statistical dynamical growth model accounting for population susceptibility. Makes predictions about the future, unconditional on particular intervention strategies.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/LANL-GrowthRate
Gao Z, Li C, Cao W, Zheng S, Bian J, Xie X, Liu TY, Zhang S, Ferres JL	Microsoft-DeepSTIA†	A deep spatio-temporal network with intervention and hospital gate under the assumption of spatio-temporal process in pandemic of different regions.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/Microsoft-DeepSTIA
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-

Lerch A, Barbera C, Poterek M, Tran Q, Moore S, Perkins A		represent the natural history of COVID-19	hub/tree/master/data-processed/NotreDame-FRED
Walraven R	RobertWalraven-ESG	Multiple skewed gaussian distribution peaks fitted to raw data.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/RobertWalraven-ESG
Nagraj VP, Turner SD, Hulme-Lowe C	SigSci_TS	Time series forecasting using ARIMA for case forecasts and lagged cases for death forecasts.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/SigSci-TS
McConnell S, Donaldson B	SteveMcConnell_ COVIDComplete	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.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/SteveMcConnell-CovidComplete https://stevemcconnell.com/covid
Bieggel H, Lega J	UA-EpiCovDA	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.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/UA-EpiCovDA
Jin X, Wang Y-X, Yan X	UCSB-ACTS	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.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/UCSB-ACTS
Wu D, Gao L, M Yian, Yu R, Vespignani A, Chinazzi M, Davis JT, Mu K, Pastore y Piontti A, Xiong X	UCSD-NEU_DeepGLEAM	Combines the signal of a discrete stochastic epidemic computational model GLEAM with a deep learning spatiotemporal forecasting framework to further improve predictions.'	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/UCSD_NEU-DeepGLEAM
Corsetti S, Schwarz T	UMich-RidgeTfReg	Nation-level model of confirmed cases and deaths	https://github.com/reichlab/covid19-forecast-

		based on ridge regression. No assumptions made about social distancing.	hub/tree/master/data-processed/UMich-RidgeTfReg
Zhang-James Y, Hess J, Chen S, Wang D, Morley CP, Faraone SV.	UpstateSU_GRU §	County-level forecast using recurrent neural network seq2seq model with the Gated recurrent units (GRU)	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/UpstateSU-GRU
Srivastava A, Prasanna VK, Xu FT	USC-SI_kJalpha §	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.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/USC-SI_kJalpha https://arxiv.org/abs/2007.05180
Srivastava A, Prasanna VK, Xu FT	USC-SI_kJalpha_RF	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	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/USC-SI_kJalpha_RF https://arxiv.org/abs/2007.05180
Woody S, et al. at the University of Texas	UT-Mobility	This model makes predictions assuming that social distancing patterns, as measured by anonymized mobile-phone GPS traces, remain constant in the future. Only models *first-wave deaths*.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/UT-Mobility
Mehrotra P, Ivan JI, and the Walmart Labs COVID-19 Team	WalmartLabsML_LogForecasting†	A logistic growth prophet forecasting model fit using case counts and deaths as features.The Model is built by Prophet model with logistic growths to forecast the US cumulative deaths. By sampling from uniform distribution to get the quantiles.	https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/WalmartLabsML-LogForecasting

* 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.