medRxiv preprint doi: https://doi.org/10.1101/2020.12.09.20246157; this version posted December 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.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-NC-ND 4.0 International license .

Assessing the Performance of COVID-19 Forecasting Models in the U.S.

Kyle J. Colonna [a*] and John S. Evans [a]

[a] Environmental Health Department, Harvard T.H. Chan School of Public Health, Harvard University, Boston, Massachusetts, USA

*** Corresponding Author:** Kyle J. Colonna (kcolonna@g.harvard.edu)

ABSTRACT

Dozens of coronavirus (COVID-19) forecasting models have been created; however, little information exists on their performance. Here we examine the performance of nine commonly-used COVID-19 forecasting models, as well as equal- and performance-weighted ensembles, based on their knowledge – i.e., accuracy and precision, and their 'self-knowledge' – i.e., 'calibration' and 'information'. Calibration and information are measures commonly employed in structured expert judgment to assess an expert's ability to meaningfully communicate the extent and limits of their knowledge.[1] Data on observed COVID-19 mortality in 4 states, selected to reflect differences in racial composition and COVID-19 case rates, over eight weeks in the summer of 2020 provided the basis for evaluating model predictions.

Only two models showed little bias (geometric mean of observed/predicted < 10%) and good precision (geometric standard deviation of observed/predicted < 1.6). Three models demonstrated good calibration and information. However, only one model exhibited superior performance in both dimensions.

Nearly all models under-predicted COVID-19 mortality, some quite substantially. Further, model performance depends on racial composition and case rates, and forecasts in the short-term outperform forecasts in the medium-term on all criteria. The performance-weighted ensembles also outperformed the equal-weighted ensemble on all criteria.

The ability of models to accurately and precisely predict mortality and the ability of the modelers to provide meaningful characterizations of the uncertainty in their estimates are potentially important to model developers and to those using model output to inform decisions.

Keywords

COVID-19; COVID-19 Decision-Making; Forecasting; Uncertainty Analysis; Cooke's Classical Method; Structured Expert Judgment

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

medRxiv preprint doi: https://doi.org/10.1101/2020.12.09.20246157; this version posted December 11, 2020. 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-NC-ND 4.0 International license .

1. Background

Effective non-pharmaceutical interventions (NPIs), or community mitigation strategies, are crucial in combating the spread of contagious illnesses like coronavirus disease 2019 (COVID-19).[2] Community NPIs, such as social distancing guidelines, restrictions, closures, and lockdowns, can effectively delay and diminish an epidemic peak, also known as flattening the epidemic curve.[2,3,4] However, these NPIs can also have immediate educational and economic consequences.[5,6,7] To make decisions on the implementation of community NPIs amidst the COVID-19 pandemic, the relevant stakeholders (e.g. government officials, community leaders, school administrators etc.) may desire estimates of the number of coronavirus cases, hospitalizations, and deaths likely to occur in their region of interest within the coming weeks.

Information about, and forecasts from, dozens of COVID-19 forecasting models are currently available via the University of Massachusetts Amherst Reich Lab's COVID-19 Forecast Hub.[8] All of the participating modeling groups provide central estimates, and most also quantitatively characterize the uncertainty in their predictions – often giving an interquartile range (25% LCL to 75% UCL) and a 90% confidence interval (5% LCL to 95% UCL) for each prediction.[9]

Unfortunately, little peer-reviewed information about the accuracy or precision of the estimates is widely available. Thus, stakeholders lack a scientific basis for deciding which models to trust and how much confidence to place in the forecasts they provide. Quite recently, one study has become available.[10] It compares the median absolute percent error of eight models stratified across world region, month of model estimation, and weeks of extrapolation.[10] Collectively, the 8 models released in July over a twelve week forecasting range had a median average percent error of about 25%, with errors tending to increase with longer forecasts and the best performing model varying by region.[10] While this information is quite useful and provides a sense of the typical bias of model predictions, other aspects of model performance may also be of interest. Users may also care about the precision of estimates and about the modeler's self-knowledge (i.e., their ability to properly characterize the uncertainty in their estimates).

Some may wonder whether better forecasts might be obtained by averaging the forecasts of two or more individual models. One such 'ensemble' model has been created by the aforementioned Reich Lab.[11] It involves an equal-weighted combination of individual model forecasts and is not performance-weighted.[11] In the expert judgment literature it has been clearly demonstrated that performance-weighted combinations of expert opinion consistently outperform equally weighted combinations.[12,13] This deserves consideration in the analysis of COVID-19 model projections.

The analysis presented below compares model forecasts with subsequent observations using several measures of model performance – reflecting 'knowledge' (bias and precision) and 'self-knowledge' (calibration and information). We also construct three ensemble models, one equal-weighted and two performance-weighted, and compare their performance with each other and with the best performing individual models. Lastly, we aim to explore whether the available models tend to provide better forecasts under certain circumstances (i.e., case rates, racial demographics, and forecast periods).

2. Methods

2.1. Data

Our analysis involves a comparison of model forecasts with subsequent observations of weekly deaths from COVID-19 in four states (Idaho, Louisiana, New York and Maine) over an eight-week period.

medRxiv preprint doi: https://doi.org/10.1101/2020.12.09.20246157; this version posted December 11, 2020. 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-NC-ND 4.0 International license .

The states considered in our analysis were selected on the basis of recent case rates of COVID-19 (cases/100,000 population within the previous week) and racial composition (majority non-Hispanic Black vs. majority non-Hispanic White). Racial composition was of interest as the COVID-19 mortality rate for non-Hispanic Black Americans is 2.1 times that of non-Hispanic White Americans.[14] With these two domains in mind, our goal was to assess two states with relatively high case rates (Idaho and Louisiana); two with relatively low case rates (Maine and New York); two with a relatively high fraction of population reported as non-Hispanic Black (Louisiana and New York); and two with a relatively high fraction of population reported as non-Hispanic White (Idaho and Maine). This was done to assess how models perform forecasting for states under varying circumstances. More detail on how the case rates and racial composition for states were determined, as well as how states were selected, is available in the supplemental material (**S.1.1.**).

We were also interested in the models' ability to forecast COVID-19 deaths in both the near-term and the medium-term. Near-term performance was gauged using projected COVID-19 deaths in the week immediately after the forecast was made. Medium-term performance was gauged using projected COVID-19 deaths in the week ending four weeks after the forecast was made.

Our evaluations of model performance for the four states and the two forecast periods of interest (week ending one week in the future and week ending four weeks in the future) were examined twice – once for forecasts made on June 13th, 2020, and a second time for forecasts made on July 11th, 2020 (no overlap in forecasts). In total, 16 comparisons of model forecasts with observed deaths were made for each model.

Of the many models providing data to the Reich Lab's data repository, only those for which all 16 forecasts were available were included in our analysis. These include: OliverWyman-Navigator (Model A)[15], MOBS-GLEAM_COVID (Model B)[16], JHU_IDD-CovidSP (Model C)[17], UMass-MechBayes (Model D)[18], UCLA-SuEIR (Model E)[19], YYG-ParamSearch (Model F)[20], UT-Mobility (Model G)[21], USACE-ERDC_SEIR (Model H)[22], and Covid19Sim-Simulator (Model I)[23].

2.2. Performance Criteria

Two aspects of model performance were evaluated – (i) the accuracy and precision of the model's central estimates; and (ii) the uncertainty of a model's estimates reflected by their reported confidence intervals.

First, to evaluate 'knowledge' (i.e., the performance of the models' *central estimates*), each observation, $O_{i,j}$, was divided by the corresponding prediction, $P_{i,j}$ – where i is an index indicating the model and j is an index reflecting the date, state, and time interval:

$$R_{i,j} = O_{i,j} / P_{i,j}$$

The distribution of the resulting ratios was then examined. For each model, i, the geometric mean (GM) and geometric standard deviation (GSD) of the distribution of $R_{i,j}$ were computed and used as measures of the observed bias and precision of the model's estimates.

Second, to evaluate 'self-knowledge' (i.e., the modelers' ability to characterize the uncertainty of their forecasts), the performance of each model was assessed using Cooke's Classical Method (CM).[1] This method was initially designed for evaluation of the performance of formally-elicited structured expert judgment (SEJ) – where an expert's ability to meaningfully characterize the uncertainty in his or her estimates is arguably as important as the predictions they provide – and has been employed in many studies.[24,25] We believe the CM is also applicable to the forecasts given by the models, as the true observations are unknown at the time of forecasting and the modeling groups may act as the 'expert' while their forecasts may serve as their 'judgment'.

medRxiv preprint doi: https://doi.org/10.1101/2020.12.09.20246157; this version posted December 11, 2020. 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-NC-ND 4.0 International license .

Cooke's CM assesses 'calibration', C, using Shannon's relative information statistic, I_s, which compares the assessed and observed probabilities of calibration variables falling within various ranges.

$$C = 1 - X_m^2 (2m * I_s), \text{ where } I_s = \Sigma_{k=1,m+1} (s_k * \ln(s_k / p_k))$$

Where $s_k = O_{i,k} / m$ is the number of realizations falling within the k^{th} of m quantiles given by the i^{th} model and p_k is the probability given by the model (the 'stated probability') that observations will fall within that quantile and X_m^2 is the Chi-squared statistic with m degrees of freedom. If s = p for all k then $I_s = 0$ and C = 1. As the divergence between stated and observed probabilities increases, I increases and C moves toward 0.

Cooke's CM assesses 'information', I, by comparing the width of the confidence intervals given by each model with the 'intrinsic range' of each calibration variable. The intrinsic range for a variable is defined as the difference between the largest forecasted or observed value and the smallest forecasted or observed value. This range is expanded slightly by multiplying it by a user-defined expansion factor, 1+ F, where F is typically a small fraction (for our data, the default 10% was used). Using this framework, the information of each expert, I_e, on each variable is defined as:

$$I_e = \Sigma_{k=1, m+1} (p_k * \ln(p_k / r_k))$$

Where p_k are the probabilities given by the model, e, and r_k are the probabilities from a uniform (or log-uniform) probability density function over the intrinsic range. Models which concentrate their forecasts in a narrow range will have high information scores.

From these two scores, Cooke calculates performance weights as the product of calibration and information and then normalizes these so that they sum to 1 across all models.

2.3. Ensemble Models

It is possible that better performance might be obtained by producing an ensemble based on weighted combinations of the forecasts given by the individual models. Two approaches of potential interest are – (i) equal-weighted combinations, and (ii) performance-weighted combinations. In addition to an equal-weighted model, two performance-weighted ensembles are considered, based on – (a) inverse-variance weights and (b) Cooke's weights.

For inverse-variance performance weighting, each random variable is weighted in inverse proportion to its variance (i.e., proportional to its precision). This weighting method is commonly used in meta-analysis.[26]

The equal-weighted model is established by equally weighting the probability densities given by the individual model forecasts. The inverse-variance-weighted model and the Cooke's-weighted model both use performance-weighted averaged densities of the individual model forecasts.

3. Results

3.1. Individual Model Performance

First, in order to visualize how accurate each model's predictions (i.e., central estimates) are, we plotted each prediction by its corresponding realization. Our results in **Figure 1** illustrate substantial differences in model accuracy.

medRxiv preprint doi: https://doi.org/10.1101/2020.12.09.20246157; this version posted December 11, 2020. 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-NC-ND 4.0 International license .

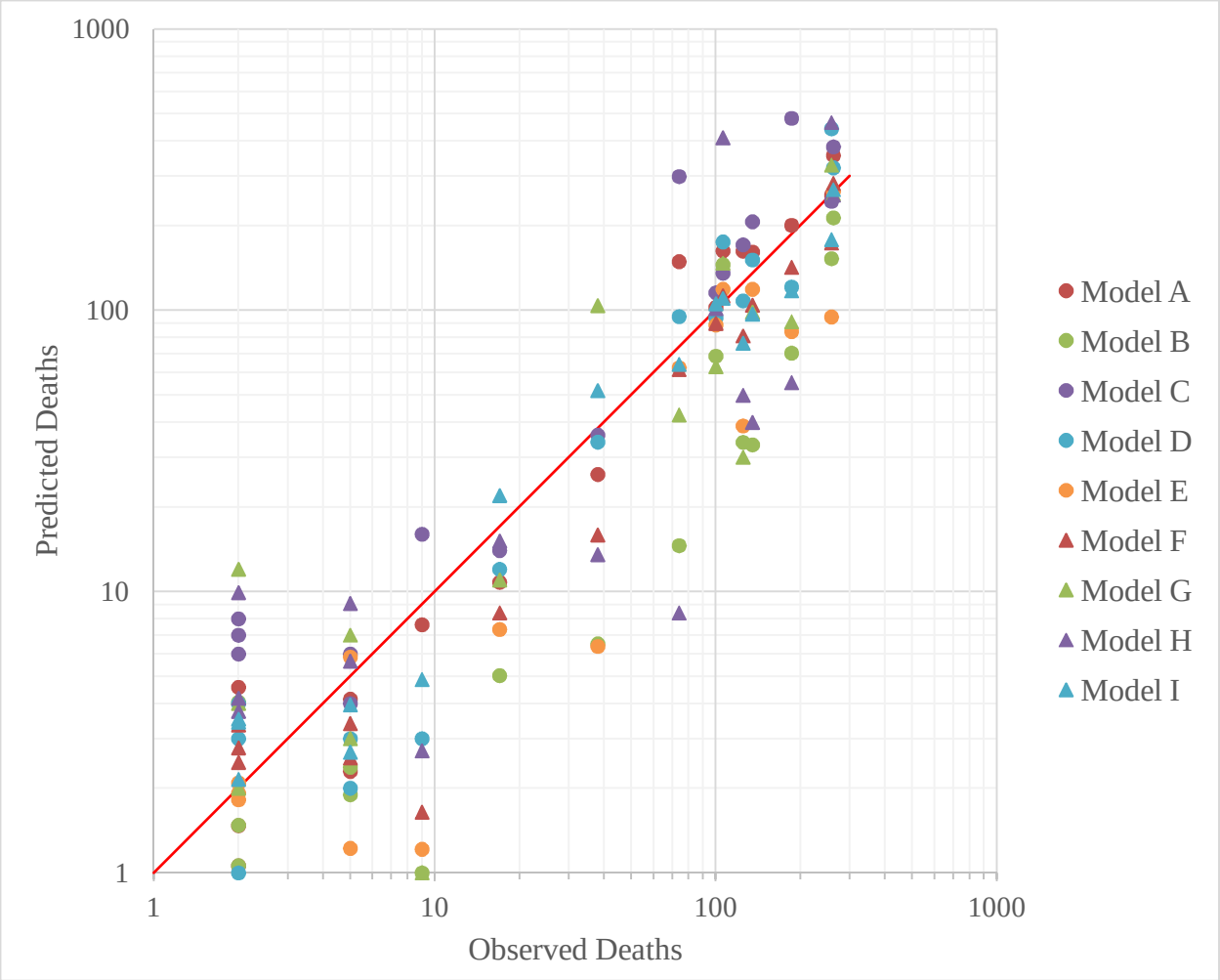

Figure 1. Observed COVID-19 mortality vs. the corresponding model central estimates in each of the four states (Idaho, Louisiana, Maine, New York) for the two forecast periods (week ending one week in the future and week ending four weeks in the future) and two forecast dates (June 13th and July 11th). The axes are on the logarithmic scale.

Each model was then assessed for all performance measures: bias (geometric mean; GM), precision (geometric standard deviation; GSD), calibration (C), information (I), and unnormalized Cooke's weight (C*I). GM and GSD are computed in regards to $R_{i,j}$. Our results for each model are summarized in **Table 1**. This table suggests that there are substantial differences in performance among these models.

medRxiv preprint doi: https://doi.org/10.1101/2020.12.09.20246157; this version posted December 11, 2020. 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-NC-ND 4.0 International license .

Table 1: Individual model performance assessed on bias (GM), precision (GSD), calibration, information, and unnormalized Cooke's weight.

Model	Bias - GM	Precision - GSD	Calibration	Information	Cooke's Weight (Unnormalized)
A	1.01	1.57	0.44	0.94	0.42
B	2.36	2.31	<< 0.01	1.68	<< 0.01
C	0.61	1.75	0.01	0.50	<< 0.01
D	1.11	1.71	0.54	1.06	0.57
E	1.78	2.04	<< 0.01	3.23	<< 0.01
F	1.35	1.70	0.29	1.38	0.40
G	1.33[i]	3.88[i]	0.04	1.41	0.06
H	1.15	2.88	<< 0.01	2.10	<< 0.01
I	1.09	1.47	<< 0.01	2.59	<< 0.01

Focusing initially on the central estimates, we see that – (i) typically, model predictions have little bias (≤ 35%), but (ii) for models B, C and E, bias is substantial. All models, except model C, have a bias score of > 1 indicating systematic underprediction of observed COVID mortality. Looking at the precision of model central estimates, we see that – (i) typically, model predictions are within a factor of ~ 2 of the observed values, but (ii) model G and H yield much less precise predictions (within factors of ~ 4 and ~ 3, respectively), and (iii) models A and I appear to offer the most precise predictions (1.57 and 1.47, respectively). Thus, if judged by the performance of their central estimates, models A and I would seem to be the most attractive – with bias < 10% and precision of < 1.6.

When we look instead at the performance of model predictions in the context of their uncertainty intervals, a different picture emerges. From this perspective, only three (or perhaps, four) of the nine models considered perform at all well – models A, D, F, and to a lesser extent, G. The calibration scores of all the other models are quite low (<< 0.01) indicating 'overconfidence' – i.e., that their stated confidence intervals are far too narrow while simultaneously poorly capturing the true value. The information scores vary from 0.5 to 3.2, suggesting substantial differences in the width of the stated confidence intervals. However, the models that have the highest information scores (models E, I, G and B) all have extremely low calibration scores (<< 0.01), suggesting that their self-confidence is misplaced.

[i] This value relies on treating model G's four-weeks ahead prediction of 0 for the week ending on July 11th in Idaho as 0.1. If instead this model G prediction is dropped from the analysis, then the GSD becomes 2.19 and the geometric mean becomes 1.00.

medRxiv preprint doi: https://doi.org/10.1101/2020.12.09.20246157; this version posted December 11, 2020. 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-NC-ND 4.0 International license .

When calibration and information are considered simultaneously (i.e., the unnormalized Cooke's weight), models D, A and F appear to perform admirably.

3.2. Ensemble Models

Each ensemble model was also assessed for all performance measures. Performance measure results for each of the ensemble models is summarized in **Table 2**.

Table 2: Ensemble model performance assessed on bias (GM), precision (GSD), calibration, information, and unnormalized Cooke's weight.

Ensemble	Bias - GM	Precision - GSD	Calibration	Information	Cooke's Weight (Unnormalized)
Equal Weight	1.24	1.60	0.03	0.72	0.02
Performance – Inverse-Variance	0.98	1.41	0.04	0.61	0.02
Performance -- Cooke's Weight [ii]	1.12	1.57	0.54	0.81	0.43

The results make clear that both performance-weighted ensembles outperform the equal-weighted combination of models. There is no measure of performance for which the equal-weighted combination outperforms both of the performance-weighted ensembles.

The inverse-variance-weighted ensemble demonstrates substantially less bias (0.98) than the equal-weighted combination (1.12). The Cooke's-weighted ensemble reflects somewhat less bias (1.12) than the equal-weighted combination but does not match the performance of the inverse-variance-weighted ensemble in this regard. In terms of precision, again the inverse-variance-weighted combination performs best (1.41), with the Cooke's-weighted ensemble (1.57) reflecting no better precision than the equal-weighted combination (1.60).

The differences in performance are more noticeable when self-knowledge (calibration, information, and Cooke's weight) is of interest. The Cooke's-weighted ensemble far outperforms the other two models in both calibration (0.54) and Cooke's weight (0.43), suggesting that it better quantifies uncertainty in its predictions (i.e., it is more accurate and provides more concentrated forecast intervals).

3.3. Performance by Domain

It is also interesting to compare these models' performance across the three domains of interest – (i) race (i.e., states which are heavily non-Hispanic White vs. states with relatively-large non-Hispanic Black populations), (ii) COVID-19 case rates (i.e., states with a relatively low amount of weekly cases per 100,000 population vs. states with a relatively high amount of weekly cases per 100,000 population), and

[ii] A hypothesis rejection significance level of 0.05 was used for this ensemble.

medRxiv preprint doi: https://doi.org/10.1101/2020.12.09.20246157; this version posted December 11, 2020. 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-NC-ND 4.0 International license .

(iii) forecast period (i.e., forecasts of mortality for the upcoming week vs. forecasts of mortality for the week ending four weeks from the date on which the forecast was made).

Table 3 evaluates the performance of equal-weighted combinations of the nine models stratified by different domains (eight forecasts per domain).

Table 3: Equal-weighted ensemble performance stratified by different domains assessed on bias (GM), precision (GSD), calibration, information, and unnormalized Cooke's weight.

Sub-Domain	Bias - GM	Precision – GSD	Calibration	Information	Cooke's Weight (Unnormalized)
High % non-Hispanic White (ID, ME)	1.30	1.86	0.60	0.86	0.52
High % non-Hispanic Black (LA, NY)	1.19	1.34	0.05	0.59	0.03
High Case Rate (ID, LA)	1.42	1.77	0.32	0.67	0.21
Low Case Rate (ME, NY)	1.09	1.39	0.14	0.78	0.11
One Week Ahead	0.95	1.41	0.27	0.53	0.14
Four Weeks Ahead	1.62	1.56	0.05	0.92	0.05

There appear to be systematic differences in the ability of these models to forecast COVID-19 mortality depending on whether they are being used to – (i) project deaths in the near or medium-term, (ii) project deaths in states which are largely non-Hispanic White or in those which have substantial non-Hispanic Black populations, and (iii) project deaths in states with low or high COVID-19 case rates.

The largest differences are seen in calibration. Model calibration is much better in states with large non-Hispanic White populations (0.60) than in states with large non-Hispanic Black populations (0.05). It is also somewhat better when making forecasts of deaths for the next week (0.27) than the week ending four

medRxiv preprint doi: https://doi.org/10.1101/2020.12.09.20246157; this version posted December 11, 2020. 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-NC-ND 4.0 International license .

weeks in the future (0.05), and in states with high case rates (0.32) than in those with low case rates (0.14).

Model forecasts are essentially unbiased (< 10% mean error) when making projections of deaths for the next week or deaths in states where case rates are low. However, these models appear to systematically underpredict deaths for the week ending four weeks in the future (GM = 1.62) and in states with high case rates (GM = 1.42). The racial composition of the state seems to have little effect on the bias of model projections.

The precision of model estimates also differs by domain, with projections being substantially less precise in both – (a) states with large non-Hispanic White populations (GSD = 1.86) than those with substantial non-Hispanic Black populations (GSD = 1.34), and (b) states with high case rates (GSD = 1.77) than in those with low case rates (GSD = 1.39). Perhaps surprisingly, there are only small differences in the precision of near-term and medium-term projections.

4. Discussion and Conclusions

Our results suggest that there may be substantive differences in the performance of the models now available to predict COVID-19 mortality in the United States and that model performance may differ when judged by 'knowledge' (i.e., their bias and precision) and self-knowledge (i.e., their calibration and information). For example, when evaluated on the basis of the bias and precision of their central estimates, models A and I would seem to be the most attractive – with bias < 10% and precision of < 1.6.

It should be noted that, although the bias of several of the models is relatively small, all but one of the models appear to systematically underpredict the true rates of COVID mortality.

When evaluated in terms of the modelers' ability to characterize the uncertainty in their estimates, a different picture emerges – models D, A and F look quite good, but model I does not. The fact that all but 3 of the 9 models considered have such low calibration scores indicates 'overconfidence' – i.e., the relatively narrow confidence intervals given by these models are unjustified and should not be relied on by users.

We also find that model performance appears to depend on the racial composition and the COVID-19 case rate for the population of interest, and that model projections of near-term mortality are better than projections of medium-term mortality.

Due to differences in inclusion criteria, only one of the 9 assessed forecasting models, YYG, was also assessed by Friedman et al. (2020).[10] Thus, it is difficult to compare performance results for the models. However, the authors do show increasing median error and median absolute error with longer forecasting periods, which is in agreement with the results shown in this analysis.[10]

Finally, our results indicate that performance-weighted ensembles outperform equal-weight ensemble models and therefore may be of interest to decision makers. The Cooke's performance-weighted ensemble outperforming the equal-weighted ensemble is in agreement with the expert judgment literature.[12,13]

Our sense is that these results are more suggestive than conclusive, because – (i) they examine only nine models, (ii) they are based on model performance in one eight-week period in the summer of 2020, (iii) they are limited to four states; (iv) they consider only forecasts of mortality and not other outcomes, such as cases or hospitalizations, that may be of interest for decision-makers; (v) our performance comparisons are descriptive and lack any formal tests of the statistical significance of observed performance

medRxiv preprint doi: https://doi.org/10.1101/2020.12.09.20246157; this version posted December 11, 2020. 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-NC-ND 4.0 International license .

differences; (vi) the performance of our ensembles has not been validated with independent data or subjected to cross-validation; and (vii) our analysis treats the models as 'black boxes' with no attempt to understand their internal structure, assumptions or data requirements.

On the other hand, our analysis has several strengths – (i) it evaluates the performance of a set of leading models which currently are being used to project COVID-19 mortality in the US; (ii) it relies on a broad set of performance criteria – which assess both knowledge (i.e., bias and precision) and self-knowledge (i.e., calibration and information); and (iii) it considers performance in four states that were selected to reflect the differences in racial composition and COVID-19 case rate in the US at the time.

medRxiv preprint doi: https://doi.org/10.1101/2020.12.09.20246157; this version posted December 11, 2020. 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-NC-ND 4.0 International license .

Availability of Data

Model forecasting data was gathered from the COVID-19 Forecast Hub's publicly available structured data storage repository on GitHub.[9] Observed state COVID-19 mortality and case data was gathered from the Centers for Disease Control and Prevention (CDC).[27] State population and racial composition data was collected from one-year estimates from the Census Bureau's 2018 American Community Survey (ACS).[28]

Table 4 in the supplemental material (**S.2.1.**) provides the racial composition statistics and case rate data. **Table 5** in the supplemental material (**S.2.2.**) provides the model predictions, their uncertainty distributions, and the subsequent observations for COVID-19 mortality. Data was analyzed using Microsoft Excel and EXCALIBUR (a software package for using Cooke's Classical Method).[29]

Acknowledgements

We want to thank Willy Aspinall, Jouni Tuomisto, and Jacqueline Macdonald for contributing to our thoughts about this and for their feedback on our early drafts of the paper. We also want to thank Roger Cooke for providing us access to and helping us to interpret results from his EXCALIBUR software.

Kyle Colonna's involvement was funded by the Harvard Population Health Sciences PhD scholarship. Prof. John Evans' involvement was funded by the Department of Environmental Health and the Harvard Cyprus Initiative at the T.F. Chan School of Public Health.

Competing Interests

The authors declare they have no competing interests that might be perceived to influence the results and/or discussion reported in this manuscript.

Author Contributions

Concept and design: All authors

Acquisition, analysis, or interpretation of data: Colonna

Drafting of the manuscript: All authors

Critical revision of the manuscript for important intellectual content: All authors

Supervision: Evans

medRxiv preprint doi: https://doi.org/10.1101/2020.12.09.20246157; this version posted December 11, 2020. 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-NC-ND 4.0 International license .

References

[1] Cooke, R. M. (1991, October 24). *Experts in Uncertainty: Opinion and Subjective Probability in Science*. Oxford University Press.

[2] DGMQ, NCEZID, CDC. (2020, April 27). *Nonpharmaceutical Interventions (NPIs)*. Centers for Disease Control and Prevention. https://www.cdc.gov/nonpharmaceutical-interventions/index.html

[3] Anderson, R. M., Heesterbeek, H., Klinkenberg, D., & Hollingsworth, T. D. (2020, March 9). How will country-based mitigation measures influence the course of the COVID-19 epidemic? *Lancet (London, England)*, *395*(10228), 931–934. https://doi.org/10.1016/S0140-6736(20)30567-5

[4] Flaxman, S., Mishra, S., Gandy, A., Unwin, H. J. T., Mellan, T. A., Coupland, H., Whittaker, C., Zhu, H., Berah, T., Eaton, J. W., Monod, M., Ghani, A. C., Donnelly, C. A., Riley, S., Vollmer, M. A. C., Ferguson, N. M., Okell, L. C., & Bhatt, S. (2020, June 8). Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature*, *584*(7820), 257–261. https://doi.org/10.1038/s41586-020-2405-7

[5] Research Dept., IMF. (2020, April 14). *World Economic Outlook, April 2020: The Great Lockdown*. International Monetary Fund. https://www.imf.org/en/Publications/WEO/Issues/2020/04/14/World-Economic-Outlook-April-2020-The-Great-Lockdown-49306

[6] *Adverse consequences of school closures*. (2020, March 10). UNESCO. https://en.unesco.org/covid19/educationresponse/consequences

[7] Nicola, M., Alsafi, Z., Sohrabi, C., Kerwan, A., Al-Jabir, A., Iosifidis, C., Agha, M., & Agha, R. (2020). The socio-economic implications of the coronavirus pandemic (COVID-19): A review. *International Journal of Surgery (London, England)*, *78*, 185–193. https://doi.org/10.1016/j.ijsu.2020.04.018

[8] Reich Lab, UMass Amherst (n.d.). *The COVID-19 Forecast Hub*. The COVID-19 Forecast Hub. Retrieved September 22, 2020, from https://covid19forecasthub.org/

[9] *Reichlab/covid19-forecast-hub*. (2020). [HTML]. The Reich Lab at UMass-Amherst. https://github.com/reichlab/covid19-forecast-hub (Original work published 2020)

[10] Friedman, J., Liu, P., Troeger, C. E., Carter, A., Reiner, R. C., Barber, R. M., Collins, J., Lim, S. S., Pigott, D. M., Vos, T., Hay, S. I., Murray, C. J. L., & Gakidou, E. (2020). Predictive performance of international COVID-19 mortality forecasting models. *MedRxiv*, 2020.07.13.20151233. https://doi.org/10.1101/2020.07.13.20151233

[11] Ray, E. L., Wattanachit, N., Niemi, J., Kanji, A. H., House, K., Cramer, E. Y., Bracher, J., Zheng, A., Yamana, T. K., Xiong, X., Woody, S., Wang, Y., Wang, L., Walraven, R. L., Tomar, V., Sherratt, K., Sheldon, D., Reiner, R. C., Prakash, B. A., … Consortium, C.-19 F. H. (2020). Ensemble Forecasts of Coronavirus Disease 2019 (COVID-19) in the U.S. *MedRxiv*, 2020.08.19.20177493. https://doi.org/10.1101/2020.08.19.20177493

[12] Colson, A. R., & Cooke, R. M. (2017, July 1). Cross validation for the classical model of structured expert judgment. *Reliability Engineering & System Safety*, *163*, 109–120. https://doi.org/10.1016/j.ress.2017.02.003

medRxiv preprint doi: https://doi.org/10.1101/2020.12.09.20246157; this version posted December 11, 2020. 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-NC-ND 4.0 International license .

---

[13] Colson, A. R., & Cooke, R. M. (2018, February 2). Expert Elicitation: Using the Classical Model to Validate Experts' Judgments. *Review of Environmental Economics and Policy*, *12*(1), 113–132. https://doi.org/10.1093/reep/rex022

[14] Division of Viral Diseases, NCIRD, CDC. (2020, August 18). *COVID-19 Hospitalization and Death by Race/Ethnicity*. Centers for Disease Control and Prevention. https://www.cdc.gov/coronavirus/2019-ncov/covid-data/investigations-discovery/hospitalization-death-by-race-ethnicity.html

[15] Oliver Wyman. *Oliver Wyman COVID-19 Pandemic Navigator*. (n.d.). Retrieved December 2, 2020, from https://pandemicnavigator.oliverwyman.com/heatmap?mode=country&region=United%20States&view=view-national&zoom=3.2&center={%22lat%22:%2038,%20%22lon%22:%20-94}&resource=resource-button-3

[16] Laboratory for the Modeling of Biological + Socio-Technical Systems. *COVID-19 MODELING*. (n.d.). Retrieved December 2, 2020, from https://covid19.gleamproject.org/mobility

[17] Infectious Disease Dynamics, JHU. *Projects | Infectious Disease Dynamics*. (n.d.). Retrieved December 2, 2020, from http://www.iddynamics.jhsph.edu/projects

[18] Sheldon, D., Gibson, C., Reich, N. (2020). *Dsheldon/covid* [Jupyter Notebook]. https://github.com/dsheldon/covid (Original work published 2020)

[19] Statistical Machine Learning Lab, UCLA. *UCLAML Combating COVID-19*. (n.d.). Retrieved December 2, 2020, from https://covid19.uclaml.org/index.html

[20] Gu, Youyang. *COVID-19 Projections Using Machine Learning*. (n.d.). COVID-19 Projections Using Machine Learning. Retrieved December 2, 2020, from https://covid19-projections.com/

[21] COVID-19 Modeling Consortium, UT Austin. *US Dashboard*. (n.d.). Retrieved December 2, 2020, from https://covid-19.tacc.utexas.edu/dashboards/us/

[22] Engineer Research and Development Center, USACE (2020). *Erdc-cv19/seir-model*. https://github.com/erdc-cv19/seir-model (Original work published 2020)

[23] MGH Institute for Technology Assessment, HMS. *Home—COVID-19 Simulator*. (n.d.). Retrieved December 2, 2020, from https://covid19sim.org/

[24] Cooke, R. M., Wilson, A. M., Tuomisto, J. T., Morales, O., Tainio, M., & Evans, J. S. (2007, September 1). A Probabilistic Characterization of the Relationship between Fine Particulate Matter and Mortality: Elicitation of European Experts. *Environmental Science & Technology*, *41*(18), 6598–6605. https://doi.org/10.1021/es0714078

[25] Hald, T., Aspinall, W., Devleesschauwer, B., Cooke, R., Corrigan, T., Havelaar, A. H., Gibb, H. J., Torgerson, P. R., Kirk, M. D., Angulo, F. J., Lake, R. J., Speybroeck, N., & Hoffmann, S. (2016, January 19). World Health Organization Estimates of the Relative Contributions of Food to the Burden of Disease Due to Selected Foodborne Hazards: A Structured Expert Elicitation. *PLoS ONE*, *11*(1). https://doi.org/10.1371/journal.pone.0145839

medRxiv preprint doi: https://doi.org/10.1101/2020.12.09.20246157; this version posted December 11, 2020. 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-NC-ND 4.0 International license .

---

[26] Hartung, J., Knapp, G., & Sinha, B. K. (2008, September 26). *Statistical Meta-Analysis with Applications*. Wiley.

[27] Surveillance Review and Response Group, CDC. (n.d.). *United States COVID-19 Cases and Deaths by State over Time*. Centers for Disease Control and Prevention. Retrieved November 24, 2020, from https://data.cdc.gov/Case-Surveillance/United-States-COVID-19-Cases-and-Deaths-by-State-o/9mfq-cb36

[28] UCSB. (2020, March 30). *American Community Survey Data*. The United States Census Bureau. https://www.census.gov/programs-surveys/acs/data.html

[29] LightTwist Software. (n.d.). *Excalibur, ExcaliburEngine*. LightTwist Software. Retrieved December 8, 2020, from https://lighttwist-software.com/excalibur/