Comparing simple and weighted ensemble forecasts of cases and deaths from COVID-19 in European countries over March to August 2021

**Background**

Short term forecasts of COVID-19 epidemiology have been a useful public health tool for understanding, communicating, and policy making in a rapidly changing epidemic. These various purposes mean that forecast accuracy is important, but also reliability across multiple forecast targets and clear and appropriate quantification of uncertainty. Meanwhile, many forecasting methods exist with varying performance under different circumstances [1].

The European COVID-19 forecast hub collates and standardises forecasts from 34 independent modelling teams for incident cases and deaths in countries across Europe. The hub has the aim of providing an accessible forecasting tool summarising forecasting efforts across the continent [#forecast-paper, #hub-repo], following similar infrastructure developed in the US [1,2] and Germany [3].

Where multiple forecasts exist for a single target outcome, the multiple predicted values can be combined into a single ensemble forecast. This creates a forecast that is typically more accurate than the individual component models [1,4] and also supports clearer communication of the most likely range of outcomes. The simplest ensemble method is to treat all contributing models equally and take an average across all forecast values. More involved methods of combination can account for the likely quality of each forecast’s predictions, by including only a selection of models, and/or adjusting each model’s forecast values by weighting based on some measure of forecast quality.

Previous work has explored the performance of ensemble forecasts from a large, standardised set of COVID-19 forecasts [5,6]. While finding that ensemble forecasts perform well, one challenge includes unstable data leading to the inclusion of outlier forecasts [7]. This is an issue where models draw from data sources subject to delays or large revisions, which can skew model parameters estimating change over time. Including these forecasts in turn distorts the performance of ensembles.

One way to address this is to simply exclude ensembles made around anomalous data points likely to have confounded the component models. However, this reduces the available pool of forecast targets with which to evaluate ensemble performance and could introduce selection bias in which ensembles are evaluated, if large data anomalies are associated with relatively more difficult (or easier) to predict patterns.

This study aims to investigate both simple and weighted ensembles. We address the problem of unstable data by using national level forecasts for weekly counts of incident cases and deaths. National data sources are more likely to be better maintained over time and less subject to large relative fluctuations around low counts [8]. Using the European COVID-19 forecasting hub should create a stable dataset of incident counts across countries which can be used to evaluate the performance of both simple and weighted ensemble forecasts.

**Methods**

*Component models*

We collected forecasts for incident weekly cases and deaths reported as COVID-19 across 32 European countries from 8 March through August 2021. Forecasts could predict between one and four weeks into the future, and predictions could be expressed deterministically (as a point forecast), or probabilistically as any subset of 23 quantile prediction intervals, with further specifications of each forecast model given in (#forecast-paper, #github-repo).

Each week we combined the available forecasts for that week into an ensemble forecast. Forecasts had to match certain criteria to be included in an ensemble. Forecasts must have been created in real time, and we did not include forecasts made after data became available. We excluded forecasts targeting or made immediately after a known anomalous data point.

*Weighting and averaging*

We combined each model’s forecasts for each target, country, horizon, and prediction interval (of 23 available quantiles) to create an ensemble. We used both mean and median methods of averaging across forecasts.

For each averaging method, we used three methods of weighting any individual predicted value. Simple ensembles took the average predicted value from all forecasts, giving equal contributions from all forecasts available for any given target. Two methods of weighting were based on past performance of an individual model, with performance taken either on average across one through four week forecast horizons, or by each week ahead horizon.

To create weights for component models, we measured past performance using the interval score. The interval score evaluates probabilistic forecasts by accounting for both calibration and sharpness of a forecast (#bracher). We excluded models which did not provide the total set of 23 prediction intervals from weighted ensembles.

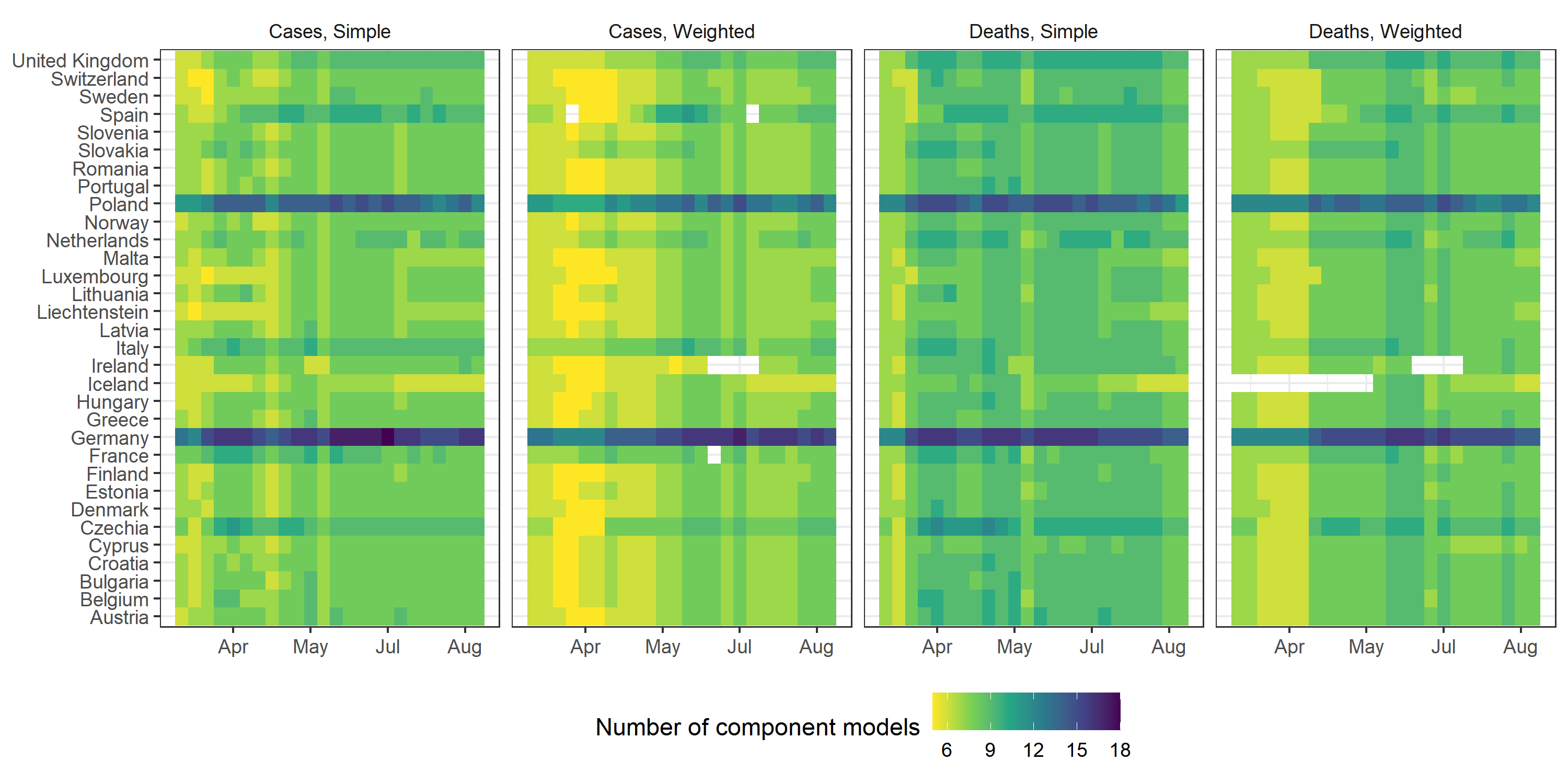
However, models varied in predicting any one or multiple targets combined from a choice of predicting case or death counts, for 32 countries, and at four forecast horizons (weeks ahead predictions). To account for this variation, we weighted the interval score based on comparing each model’s score to every other model forecasting for the same target, creating a pairwise comparison tournament. We then took the geometric mean of these pairwise comparisons for each model. This resulted in a single score per model for each of two target counts, 32 locations, and four forecast horizons. Separately, at this point we also averaged these scores across forecast horizons.

To evaluate these scores against the simplest possible prediction, we set a baseline model of a flat forecast with expanding bounds of uncertainty. We took the weighted interval score of each model and scaled it against the performance of the baseline forecast, giving a measure of performance that accounted for each forecast’s individual skill compared to all other equivalent forecasts and a simple baseline. We took the inverse of these scores to create weights on a scale of 0-1 and applied these to a model’s forecast values at all quantile predictions for each model. We then averaged across these weighted values at each quantile.

To evaluate the simple and weighted mean and median ensemble forecasts, we used the same measure of performance described above based on calculating the relative interval score scaled to a baseline. We also explored measures of coverage, the extent to which a probabilistic prediction was calibrated to observations and compared the performance of each ensemble across each target count, location, and forecasting horizon.

**Results**

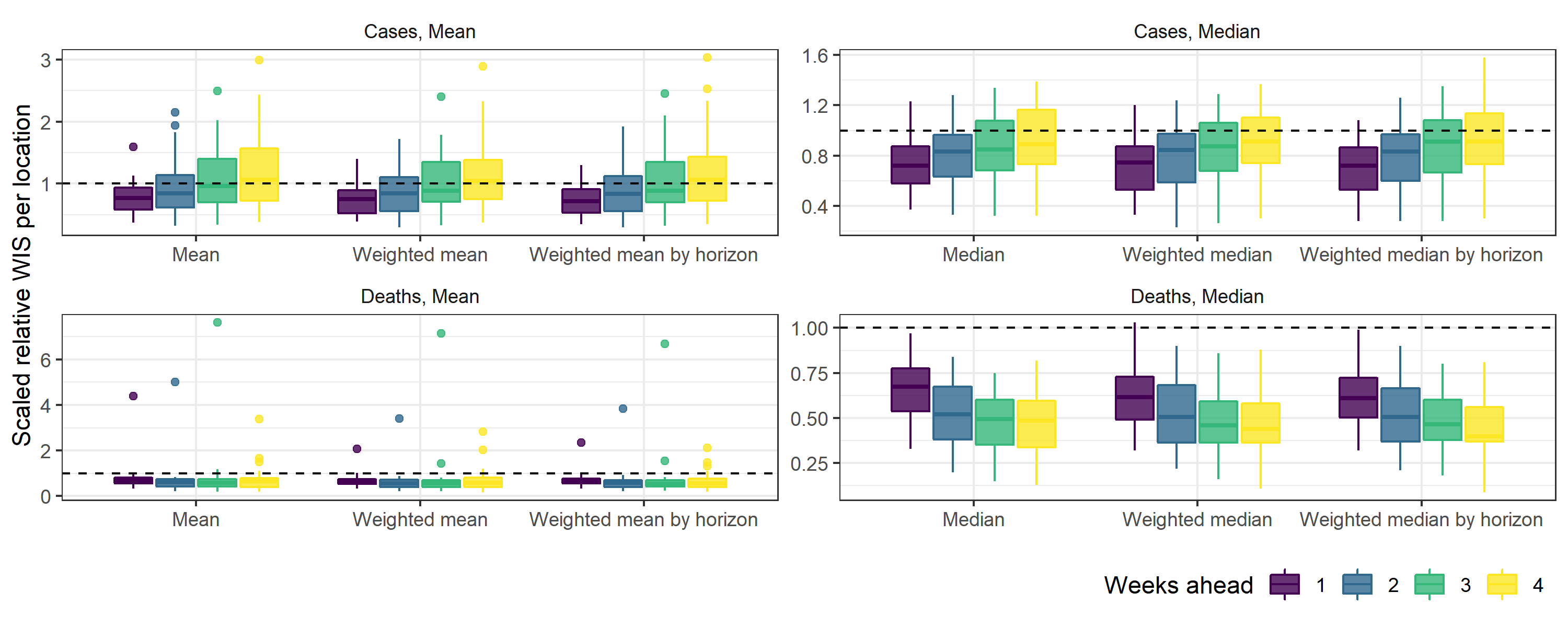
*Component models*



*Figure #1: Number of component forecasts over time for each country, by target count and ensemble method (simple or weighted).*

We created forecasts over the period 2021-03-15 to 2021-08-23. For each week, we created six probabilistic ensemble forecasts of incident weekly case and death counts for 32 countries, forecasting over one through to four weeks (a combined 256 targets). This created 1532 evaluated ensembles, after removing 4 weeks of forecasts in countries reporting data anomalies.

We collected forecasts from a total 29 modelling teams. Ensembles taking the simple averages (mean, median) of all forecasts for each week included between 5 and 18 component models over time. The weighted average ensembles used a stricter set of inclusion criteria, reducing the number of component models to between 5 and 17.

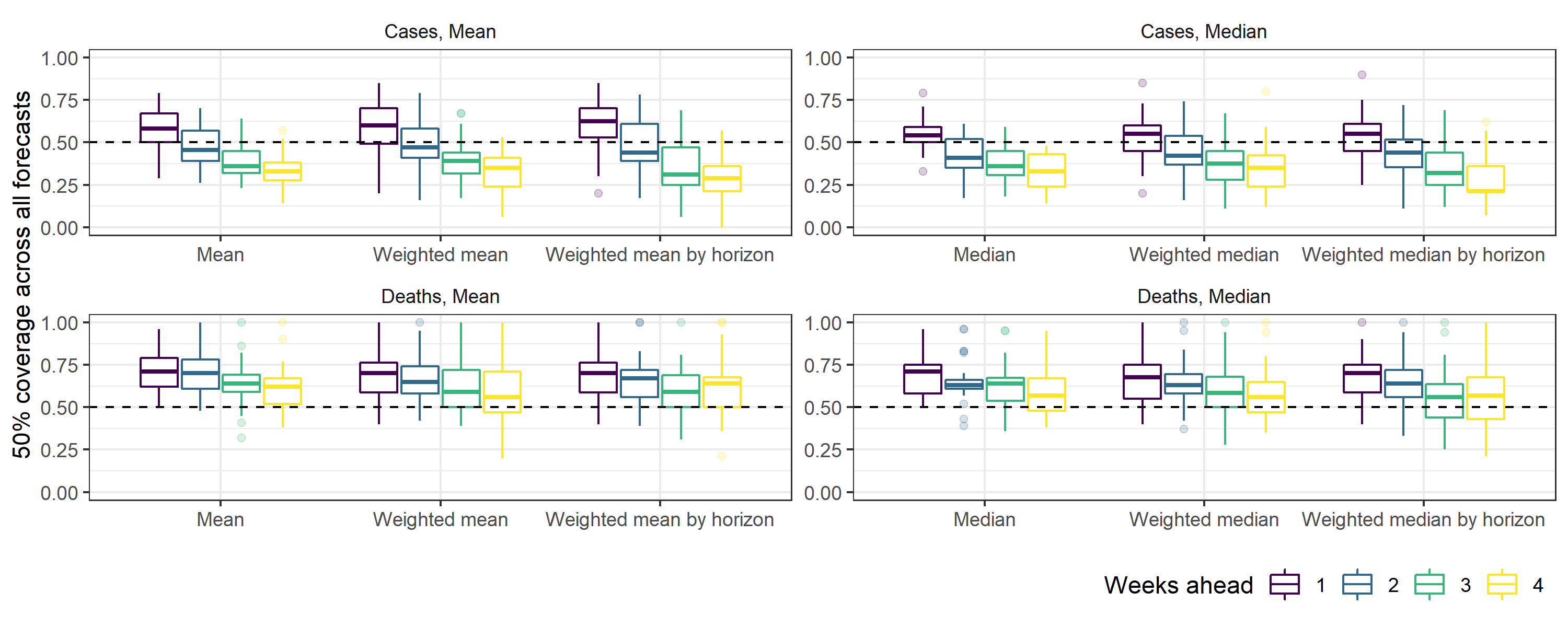


*Figure #2: Relative interval score compared to baseline forecast model (dashed line, 1), by ensemble method (mean or median, with weighted and weighted by horizon methods), target count (cases or deaths), and weeks ahead horizon (1 through 4). Boxplots show distribution of scores across locations, with anomalous ensemble performance (countries with much better or worse forecasts than average) shown as points.*

*Relative performance and uncertainty*

For the majority of forecast targets, ensembles performed better than the baseline. Of the 1532 combinations of the six ensemble methods each forecasting four weeks of case and death counts in 32 countries, 82% performed better than the baseline forecast model. Ensemble forecasts consistently performed better compared to the baseline model when forecasting incident deaths. With a total 766 targets for incident deaths, 96% ensembles outperformed the baseline model. This was 67% of the same number of case targets.

For all ensemble methods, the skill of ensemble forecasts varied from the near to further into the future, relative to the baseline model. For each model this was always a consistent trend in gradually improving or worsening skill over longer horizons. However, the direction of trend varied by the epidemiological target being forecast. Forecasting cases proved more difficult at longer forecast targets than shorter horizons, while the relative skill of ensembles in forecasting incident deaths improved over longer horizons.



*Figure #3: Coverage: The proportion of observations that fell within the 50% prediction interval for each ensemble, by target count of cases and deaths and horizon. Ideally, a forecast model would achieve 50% coverage of 0.50 (meaning 50% of observations fall within the 50% prediction interval), shown as the vertical dotted line. Values of greater than 0.5 indicate that the forecasts are under-confident (prediction intervals are on average too wide), whereas values smaller than 0.5 indicate that the forecasts are overconfident (prediction intervals tend to be too narrow.)*

When forecasting deaths, ensembles of any model across all horizons typically gave too broad a range of uncertainty (figure #3). This was also true when forecasting case counts at the one-week ahead horizon for both mean and median varieties of ensemble.

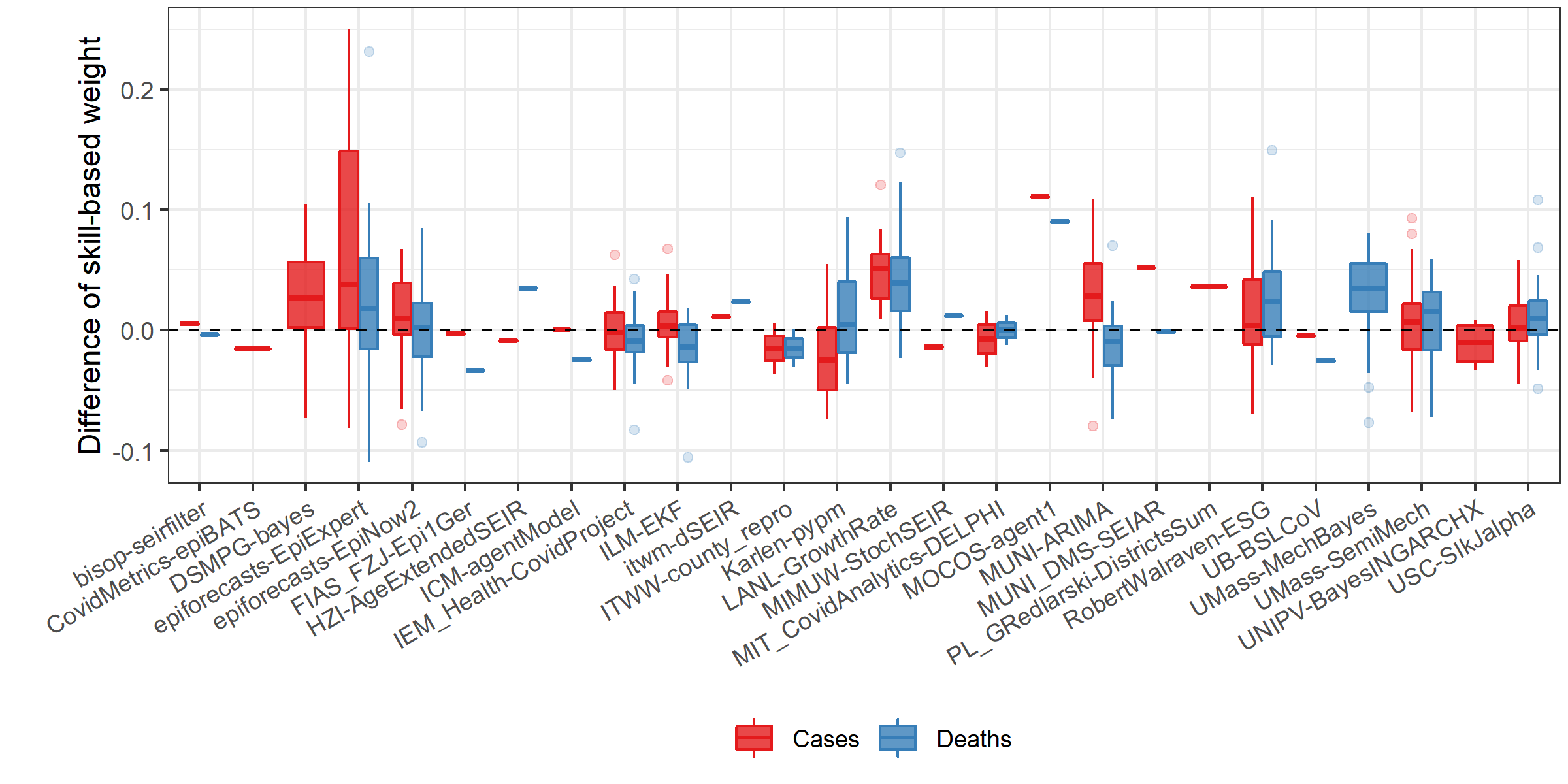
For both targets, the problem of underconfident prediction intervals reduced over longer horizons. This only improved forecasts of incident deaths, with the coverage of the 50% prediction interval the most accurate at four weeks. However, case forecasts became overconfident in nearly all locations by the three and four week horizons.

*Averaging methods*

Ensemble forecasts that used any form of median outperformed the baseline model across all horizons and for each of 32 locations. The 766 ensembles using a median outperformed the baseline for 85% targets. Ensembles using the mean were less consistent in performance across countries, while still outperforming the baseline for 79% targets.

Forecasts for Iceland were a notable outlier, where mean forecasts failed to accurately forecast regardless of the type of mean ensemble used. This includes the worst of any ensemble performance, where the simple mean of forecasts for incident deaths in Iceland at three weeks ahead performed over seven times worse than the baseline forecast.

*Weighting methods*



*Figure #4: Change in model contribution of each model to ensemble forecasts: difference from weighting in a simple ensemble (dashed line, 0), to weighted by relative skill (averaged across all horizons), by target. Boxplot distribution represents average difference in model weight for each country, where data above the dashed line indicates that the model weight increased, and with anomalous differences shown as points.*

Weighting by skill created substantial differences in the contribution of individual models compared to weighting all models equally (figure #4), and these differences varied by target and over time (for more on relative performance of individual models, see forecast report). However, weighting by skill had relatively little impact on the performance of ensembles compared to simple equal weights. Of 256 targets, there was no greater than 2% difference in the number of ensemble forecasts that outperformed the baseline, between ensembles weighted by skill on average, by horizon, or a simple average for each of the mean and median approaches.

**Discussion**

We explored the combination of individual forecasts into an ensemble forecast for incident cases and deaths due to COVID-19 across 32 European countries over March to August 2021. Of 1536 possible ensemble forecasts we only excluded four forecast targets, indicating the value of evaluating against a cross-country national level dataset. We used variations on methods for weighting the ensemble (as equal weights or weighted by component forecasts’ relative skill), and for combining into an average (as mean or median average across each quantile prediction). We found that ensemble forecasts overall performed better than a simple baseline forecast.

Ensemble forecasts for incident death counts had consistently better performance than forecasts of case counts. We also noted that performance improved at longer horizons for death forecasts of any ensemble method, while performance worsened over time for case forecasts. This suggests that even while ensemble forecasts at short time horizons are useful for predicting both cases and deaths, the greatest value of ensemble forecasting is in providing the most accurate forecast for deaths over long time periods.

Using any form of median forecast was nearly always the most accurate choice of ensemble method. This compares to previous ensembles of standardised epidemiological forecasts from individual modelling teams. Ray and others explored three methods to combine forecasts of COVID-19 reported deaths in US states [6]. Using a simple mean, simple median, trained (weighted) mean (over the full history and over a 4 week window), they found the trained mean performs similarly to the simple median, and both outperformed the simple mean.

Ensembles using weights based on past performance may be biased depending on the assessment of past performance. This is likely if performance is compared between models forecasting for different targets. To avoid this, we used the interval score, combining forecast sharpness in a pairwise comparison, further scaled against a flat-line baseline forecast. This meant that models were compared on equivalent targets and against a clear benchmark.

One area of bias in model weights could be influenced by epidemic dynamics in Europe over the study period of March to August 2021. Forecast models of incident cases had to adapt where the B.1.617.2 (delta) variant overtook the existing SARS-CoV-2 viral strain in driving new cases, with likely increased transmissibility and a shorter generation time [9]. Meanwhile, the introduction of mass vaccination likely influenced models of incident deaths, particularly where varying population vaccination strategies meant the effect of a rise in cases was less consistent across countries. Together this may have meant that the ensemble weights produced by evaluating models’ past performance over all time were poorly calibrated to forecast these changing dynamics. One area for further work can therefore be to compare an ensemble with weights calculated by only more recent past performance, or performance divided into epidemic phase.

This supports the use of an ensemble as both a useful as well as an accessible way to present forecasts, summarising across a large combination of both diverse models and forecasting targets. There was no clear best performer among the individual models that contributed to the ensembles, for any specific target (#forecast-paper). Based on our findings, we recommend the use of median ensemble methods and that policy relevant work that uses ensembles should place more confidence in forecasts of incident death than case counts, particularly at longer (3-4 week) periods into the future.

**References**

[1] Reich NG, Brooks LC, Fox SJ, Kandula S, McGowan CJ, Moore E, et al. A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States. Proc Natl Acad Sci 2019;116:3146–54. https://doi.org/10.1073/pnas.1812594116.

[2] Cramer EY, Ray EL, Lopez VK, Bracher J, Brennen A, Rivadeneira AJC, et al. Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the US. MedRxiv 2021:2021.02.03.21250974. https://doi.org/10.1101/2021.02.03.21250974.

[3] Bracher J, Wolffram D, Deuschel J, Görgen K, Ketterer JL, Ullrich A, et al. Short-term forecasting of COVID-19 in Germany and Poland during the second wave – a preregistered study. 2021. https://doi.org/10.1101/2020.12.24.20248826.

[4] Brooks L. Comparing ensemble approaches for short-term probabilistic COVID-19 forecasts in the U.S. - International Institute of Forecasters 2020. https://forecasters.org/blog/2020/10/28/comparing-ensemble-approaches-for-short-term-probabilistic-covid-19-forecasts-in-the-u-s/ (accessed July 15, 2021).

[5] Taylor JW, Taylor KS. Combining Probabilistic Forecasts of COVID-19 Mortality in the United States. Eur J Oper Res 2021. https://doi.org/10.1016/j.ejor.2021.06.044.

[6] 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. https://doi.org/10.1101/2020.08.19.20177493.

[7] Ray E. Challenges in training ensembles to forecast COVID-19 cases and deaths in the United States - International Institute of Forecasters 2021. https://forecasters.org/blog/2021/04/09/challenges-in-training-ensembles-to-forecast-covid-19-cases-and-deaths-in-the-united-states/ (accessed August 5, 2021).

[8] Moran KR, Fairchild G, Generous N, Hickmann K, Osthus D, Priedhorsky R, et al. Epidemic Forecasting is Messier Than Weather Forecasting: The Role of Human Behavior and Internet Data Streams in Epidemic Forecast. J Infect Dis 2016;214:S404–8. https://doi.org/10.1093/infdis/jiw375.

[9] Lopez Bernal J, Andrews N, Gower C, Gallagher E, Simmons R, Thelwall S, et al. Effectiveness of Covid-19 Vaccines against the B.1.617.2 (Delta) Variant. N Engl J Med 2021;385:585–94. https://doi.org/10.1056/NEJMoa2108891.