Results

K Sherratt

28/08/2021

**Results**

Over March through August 2021, epidemic dynamics diverged between countries covered by the European forecast hub. Here we focus on the development and performance of contributions to the forecast hub, while more general description of the spatio-temporal distribution of COVID-19 cases and deaths is readily available from the ECDC (#link-ecdc-dashboard).

*Forecast community*

We actively recruited forecasters over February and March 2021 and accepted forecasts each Monday from 8 March 2021. Between 19 and 31 teams independent from the Hub then submitted forecasts each week, with a total of 34 independent teams contributing at least one forecast over the period to 23 August 2021. Forecasting teams contributed from institutions across 7 European countries, with 10 teams’ institutions based outside the geographical scope of the Hub (in Canada, Australia, and the USA).

We received 25 free-text responses to the question “What is your main motivation for submitting (or not submitting) to the forecast hub?”. 19 responses contained motivations in favour of submitting, from which we identified two themes positively motivating participation. Respondents valued the opportunity to contribute to the ensemble specifically, to the wider scientific aims of the project, and to gain policy visibility for forecast results. Respondents also used the hub to develop their own modelling methods, through learning from meetings, comparison between their and others’ models, or model evaluation provided by the hub. Seven responses named barriers to submitting, including forecasting for different targets or time periods than that used by the hub, and time constraints.

We held a total of 19 hour-long weekly meetings between 9 March and 20 July. Persistent discussion themes included handling the spread of new variant strains including Alpha and Delta strains, and variable uptake and impact of vaccines. There was also some evidence of public engagement in the Hub. Figures from the ensemble were quoted in news articles (#news-3-examples), and results from the ensemble forecast were directly quoted in weekly ECDC surveillance report aimed at public health practitioners (#ecdc-surveillance). To support the creation of new epidemiological forecast hubs, we documented our adaptations from the US and German-Poland hubs and contribute to a shared R package for working with Hub forecasts (#covidHubUtils). We are developing a generalised code base which can be easily used to create new forecast hubs for any targets or locations.

*Forecast models*

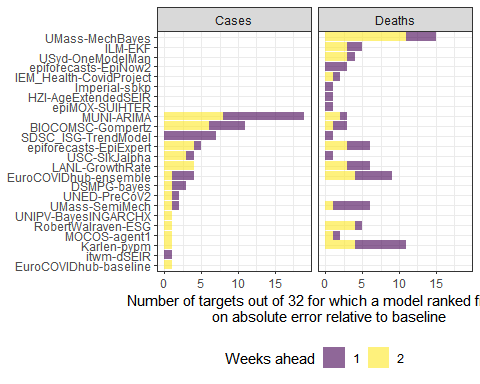
Among the 34 teams submitting to the hub, 2 teams contributed forecast results from multiple separate models. We added a simple baseline forecast and an ensemble of all submitted forecasts, resulting in a total models\_unique separate forecasting models. Among these, 61% (25 models) forecast for multiple countries, including 16 models with forecasts for all 32 hub locations. 73% (models\_all\_targets\_n models) submitted forecasts for both incident cases and incident deaths, with a further 6 models forecasting only cases and 5 forecasting only incident death counts. All models forecast for targets at one week ahead, with most (85%, 35 models) also forecasting for two weeks ahead. Over 80% of models also forecast for three and four week ahead targets.

Forecast methods ranged from data-based machine learning and statistical models to theory-driven structured models, as well as blends of different approaches. Teams also varied in their approach to quantifying uncertainty. Nearly all (95%, 39 models) gave probabilistic forecasts. 37 models always included predictions across the entire pre-specified set of 23 quantiles, while a further 5 models expanded to include the full range of uncertainty after initially submitting only the median (point) or a subset of quantile predictions. 2 models submitted only a subset of quantiles, and a further 2 models submitted point forecasts, with the single forecast value assumed to represent the median quantile for each target.

*Forecast performance*

We evaluated predicted counts against JHU data and excluded forecasts around data anomalies in France on the #, Spain on the #, and Ireland over the period # to #. We compared model performance relative to a baseline of the simplest possible probabilistic forecast and observed that forecasting models displayed wide variability in relative performance forecasting for different locations and for cases and deaths. We focus on evaluating forecasts made for one and two week horizons and assess the accuracy and consistency of forecast models across countries as well as their coverage of uncertainty.

**- Evaluation is based on all models that have submitted > 4 forecasts over all time. This is different to the latest evaluation csv which excludes forecasts which haven’t submitted in the prior two weeks. All the code and reporting works with both, so this report can be re-generated for either evaluation option. It changes the total number of forecasts and relative scores change a little but the surrounding text stays the same**

*1. Accuracy and consistency* 

*Figure #: Average of scaled relative absolute error in forecasts of one and two week ahead incident cases and deaths, by model. The vertical dashed line represents the baseline to which scores are scaled. The score is the absolute error, the difference between the median forecast prediction and the observed value for each forecasting week; this was scaled against the score of a simple baseline forecast and averaged over the number of weeks the model forecast for each target country. Single points indicate model forecasts for a single location, boxplots indicate the distribution of model performance where models forecast across locations.*

*Figure #: Models by number of forecast targets where each was the top ranked model. Forecast targets include 128 combinations of cases/deaths, one and two week ahead horizons, and 32 hub locations; total N = 152 including ties for top rank.*

To compare accuracy and consistency across both probabilistic and deterministic forecasts, we focus on the mean average of the absolute error relative to the baseline, where lower scores relative to the baseline are better (figure #). Across 128 separate targets (cases or deaths at one or two weeks ahead among 32 locations), a variety of individual models consistently outperformed the baseline for single targets (figure #). Only nineteen models ranked first for any target among the 40 models (all those submitted plus the hub ensemble), with one model leading performance on nearly 20% of forecast targets. However, the hub ensemble was the only model which combined accuracy with consistency across targets. The ensemble was the only model that ranked among the top three models on mean average score relative to baseline for both cases and deaths and at both one and two week ahead forecasts. When forecasting one week ahead counts, the ensemble performed on average worse than the baseline in only four out of 32 countries for both cases and deaths. It also ranked first or second among the narrowest range of scores, with a narrower range indicating more consistent performance, with the exception only of forecasting cases at a two week horizon.

*Figure #: One week ahead average forecast scores across all models for each country, against the score of a baseline model for that country, for cases and deaths. The averaged forecast score is either the interval score for probabilistic forecasts or the mean absolute error for single point forecasts.*

In terms of accuracy, the average performance of models in predicting cases tracked near to the baseline forecast across nearly all countries (figure #). For one week ahead cases, Poland, Romania, and Austria saw the best performances on average relative to the baseline, with a scaled score average of 0.61, 0.65, and 0.67 respectively. In contrast, models were in general better than the baseline at predicting deaths, with the average of scaled model scores beating the baseline in 20 and 21 of 32 countries at one and two week horizons respectively. Meanwhile, unusual epidemic dynamics could cause single models to heavily skew these averages by country. The average score of forecasts for deaths in Iceland was heavily skewed by a single model forecasting a sudden increase in deaths, which, despite an outbreak of cases, remained at no loss over the entire period.

Teams took a range of epidemiological modelling approaches to forecasting individual country dynamics in cases and deaths, at a time when epidemic patterns started to diverge among countries due to increasingly varying social restrictions, vaccination rates, and variant emergence. This may have been reflected in forecast scores. We briefly explored the ability of models to generalise across targets by country and variable. Of ten models that forecast one-week ahead cases for only a single country, nine performed better than the baseline model. This compares to the nine models that forecast one week ahead cases for all 32 locations. Among these 288 combined forecasts, only 52% (149 forecasts) outperformed the baseline. By contrast, these proportions were relatively similar for death forecasts. For eight single-country models forecasting one week ahead deaths, five (63%) outperformed the baseline, while for the nine models covering 32 countries, 65% (186) of forecasts outperformed.

*Figure #: One week ahead forecasts and observed counts for weekly incident cases in France. Points represent individual model point (median) predictions each week, with dotted lines as 25% and 75% quantile predictions for those models that provided them. The black line represents observed count. Forecasts and data excluded around a data anomaly (-350,000 cases reported) in May.*

We observed that forecasts for counts in some target locations saw high errors across nearly all models. In France, despite excluding forecasts around a major data anomaly, all forecasts for one week ahead incident cases saw higher absolute errors than any other location (figures #, #). Across non-hub forecasts, this target had an average 28,080 mean error (average of interval score or absolute error for point forecasts). This compared to an average error across all locations at 3,952. However, this can be used to highlight models that performed particularly well against a generally difficult to predict outcome. The MUNI-ARIMA model had a significantly lower error (WIS 12,802) when forecasting one week ahead cases in France and was the only individual model other than the hub ensemble to out-perform the baseline (WIS 24,212) for this target.

*2. Coverage of uncertainty*

*Figure #: The proportion of observations that fell within the 50% prediction interval for each model, by target count of cases and deaths at a one and two week 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.) The mean and 95% confidence interval are calculated across coverage scores by location.*

We explored the calibration of uncertainty for 31 models which provided predictions for each target across the full quantile distribution (figure #). Summarised on average across all models, calibration of uncertainty was near perfect. Across all models, coverage of the observed 50% distribution matched near-perfectly, at 50% for one-week ahead forecasts of deaths and 48% for similar coverage of cases. However, between separate models, average coverage ranged from 4% to 95% of the observed 50% interval for deaths and 20-89% for cases. This in turn meant that despite taking all models into an average forecast, the hub ensemble also suffered with imperfect coverage. For one week ahead forecasts, the ensemble’s calibration to the 50% distribution of observations was at 57% for cases and 71% for deaths.

Compared to deaths, forecasts for cases had a more stable range of coverage at the 50% observed distribution. Of 23 models that gave probabilistic forecasts for cases one week ahead, around half (12) averaged between 40% and 60% in attempting to forecast the 50% observed distribution. However, seven models were generally overconfident in their central forecasts (predictions in the 50% interval matched less than 40% observations), while four were underconfident. Death forecasts were more biased to over-confidence, with 10 of 24 models forecasting less than 40% of the 50% distribution at one week ahead.

Coverage worsened slightly at longer horizons (averaging 41% and 51% for two-week case and death forecasts respectively). There was no clear pattern of coverage level that remained consistent between the two targets of cases and deaths. At a one week horizon, 10 out of 19 models that forecast for both targets had a consistent pattern of under or over coverage, while the remaining nine under-predicted for cases and over-predicted for deaths or vice versa. Meanwhile, there was relatively little variation in coverage between forecasts for different locations from the same model (figure #).