Analysis UK Crowd Forecasting Challenge

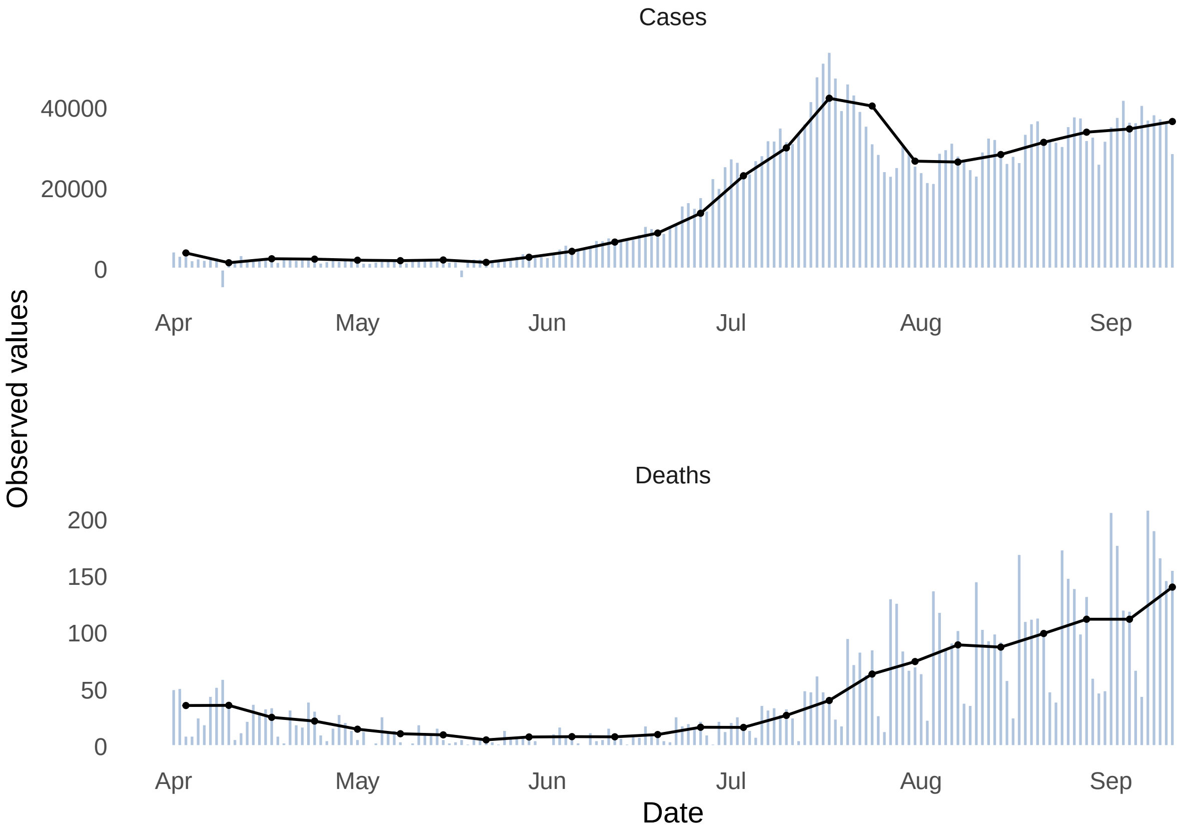
This is an analysis of forecasts made by participants of the UK COVID-19 Crowd Forecasting Challenge. Over the course of 13 weeks (from May 24 2021 to August 16 2021) participants submitted forecasts using the crowdforecastr prediction platform.

These forecasts were aggregated by calculating the median prediction of all forecasts. These aggregated forecasts (later denoted as "epiforecasts-EpiExpert" or "Median ensemble") were submitted to the European Forecast Hub.

1 Prediction targets and observed values

Participants were asked to make one to four week ahead predictions of the weekly number of reported cases and deaths from COVID-19 in the UK.

Here is a visualisation of daily (bars) and weekly (line) reported numbers.



2 Forecast evaluation

Forecasts were evaluated using the "weighted interval score" (WIS). This 'score' is negatively oriented, meaning that a lower score is better. You can think of the weighted interval score as a 'penalty' for being less than perfect.

Ranking	Forecaster	Score
1	anonymous_Stingray	4.941835
2	seb	6.412597
3	aen	6.561190
4	Trebuchet01	6.677199
5	habakuk (Rt)	6.704598
6	Gw3n	6.753411
7	aurelwu	6.774893
8	Cantabulous	6.781753
9	seb (Rt)	6.807969
10	olane (Rt)	6.820095

The weighted interval score is the sum of three components (i.e. three different types of penalties): “overprediction”, “underprediction” and “dispersion”. Overprediction and underprediction are penalties that occur if the true observed value falls outside of the range of values deemed plausible by a forecast. If a forecast is very uncertain, then the range of plausible values is larger and it is less likely to get penalties for over- and underprediction. The “dispersion” term on the other hand penalises a forecast for being overly uncertain.

To make forecasts of deaths and reported infections more comparable, we took the logarithm of all forecasts as well as the logarithm of the “ground truth data” and then calculated the weighted interval score using these.

This is different from the methodology used by the European Forecast Hub, which does not take the logarithm of forecasts and observed values. Taking the logarithm means that forecasts are scored in relative terms rather than absolute terms. On the natural scale it is important whether a forecast is e.g. 10 off or 1000, while on the logarithmic scale we score whether a forecast is 5% or 10% off - regardless of the absolute values. This may make more sense for a pandemic anyway where infections spread exponentially. It also allowed us to combine death forecasts and case forecasts and compute a single score to rank forecasters.

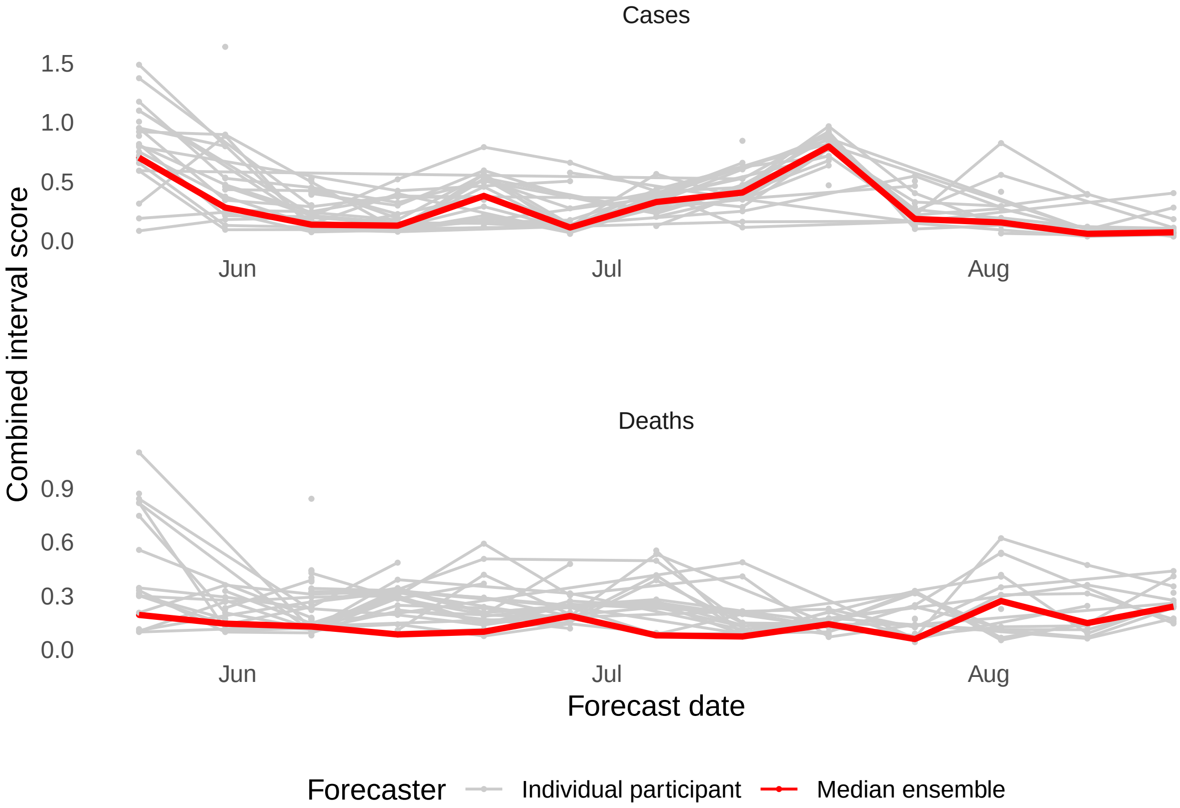
If a forecaster did not submit a forecast for a given forecast date, they were assigned the median score of all participants who submitted a forecast on that day.

2.1 Leaderboard

Here is the official leaderboard with overall performance summarised over all forecasts

2.2 Individual vs. ensemble performance over time

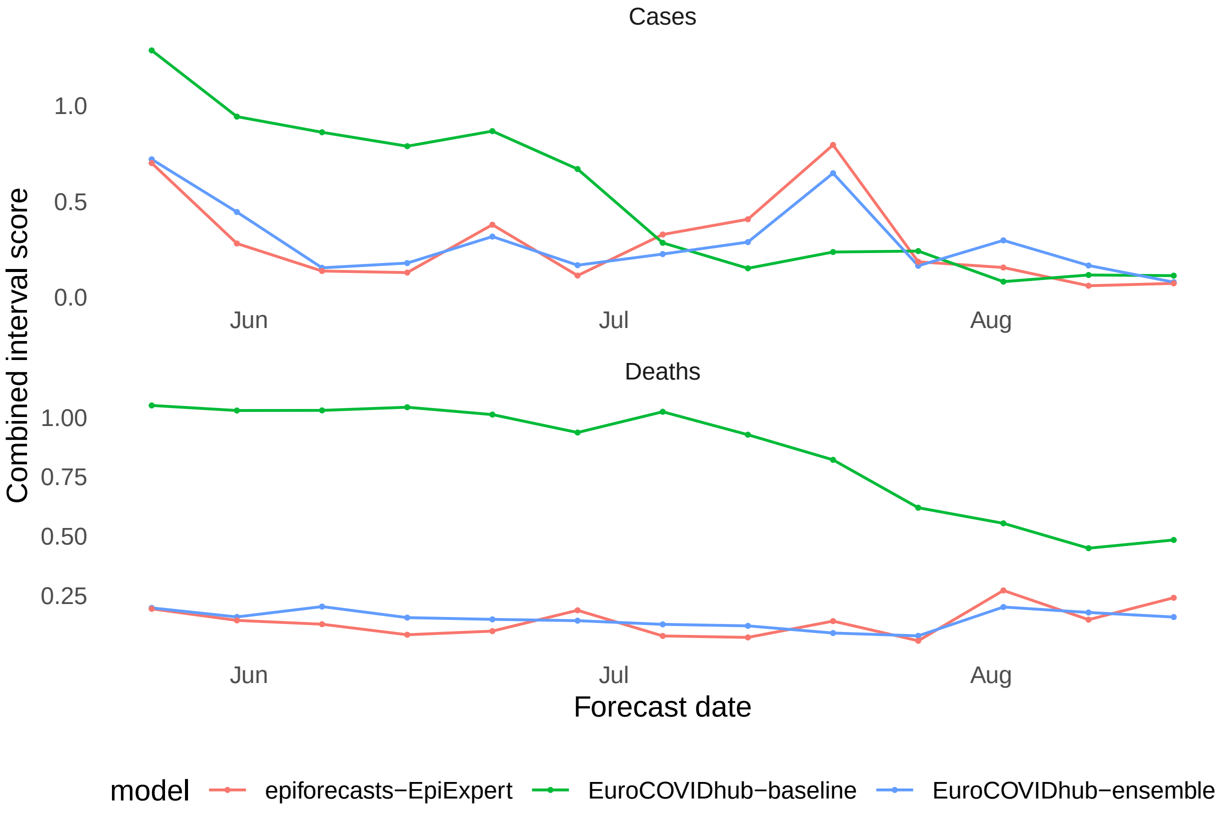
Here is a visualisation of individual participants’ scores together with scores from the median ensemble of all forecasts shown in red.



2.3 Comparison against the Forecast Hub

This visualisation compares the ensemble of all forecasts from participants in the UK Crowd Forecasting Challenge ("epiforecasts-EpiExpert) against the ensemble of all forecasts from the European Forecast Hub (including our own forecasts) and the Forecast Hub baseline model.

model	target_type	score
epiforecasts-EpiExpert	Cases	3.716544
EuroCOVIDhub-ensemble	Cases	3.822955
EuroCOVIDhub-baseline	Cases	6.622201
epiforecasts-EpiExpert	Deaths	1.841795
EuroCOVIDhub-ensemble	Deaths	1.959891
EuroCOVIDhub-baseline	Deaths	10.969080



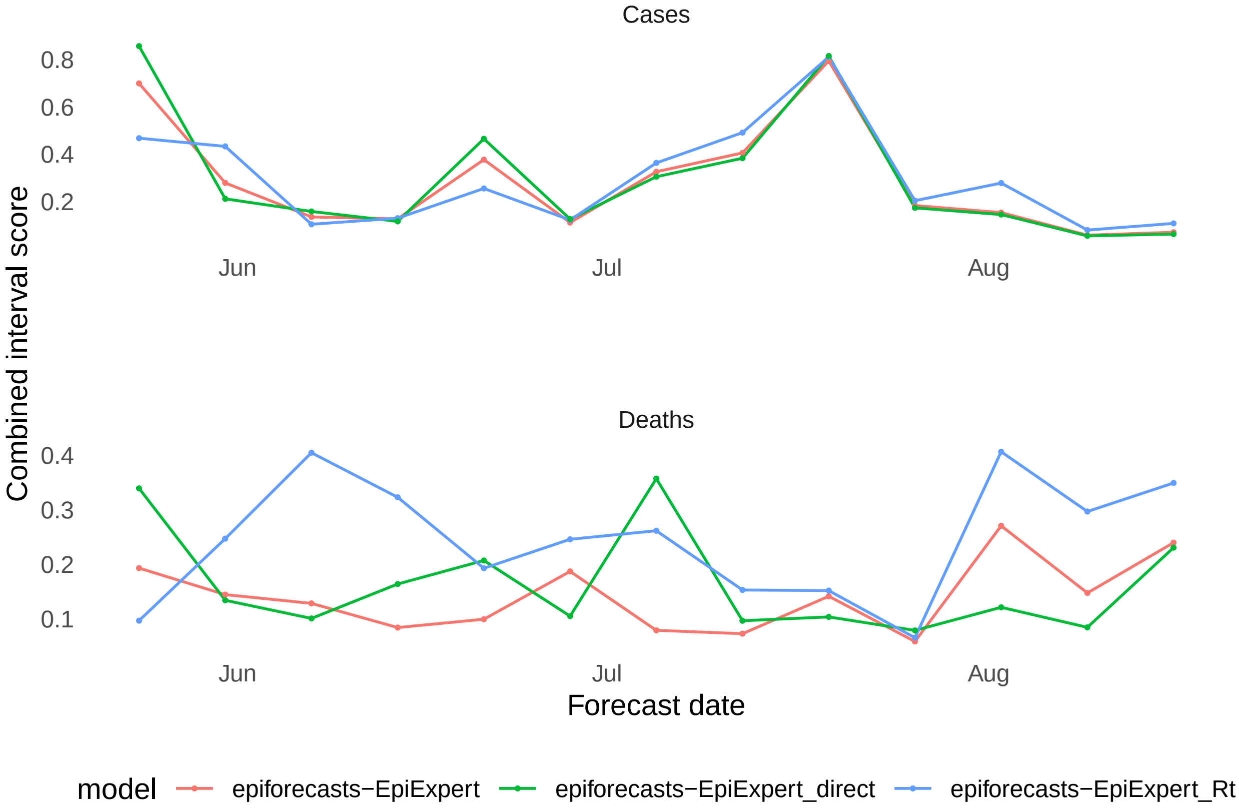
Here is a summary of scores:

2.4 Comparison of the different EpiExpert forecasts

Users could submit two different forecasts. One was a direct forecast of cases and deaths. The median ensemble of these direct forecasts is called "EpiExpert_direct". The other one was a forecast of R_t, the effective reproduction number. This R_t forecast was then mapped to cases and deaths using the so-called renewal equation, which models future cases as a weighted sum of past cases times R_t. The median ensemble that uses only these forecasts is called "EpiExpert_Rt". The "EpiExpert" ensemble is a median ensemble that used both regular as well as R_t forecasts.

Here is a visualisation over time:

model	target_type	score
epiforecasts-EpiExpert	Cases	3.716544
epiforecasts-EpiExpert_Rt	Cases	3.845169
epiforecasts-EpiExpert_direct	Cases	3.870579
epiforecasts-EpiExpert	Deaths	1.841795
epiforecasts-EpiExpert_direct	Deaths	2.118421
epiforecasts-EpiExpert_Rt	Deaths	3.190695



And a summary of scores:

3 Number of available forecasts

The following gives an overview of the number of forecasts made

3.1 Number of forecasts per forecaster

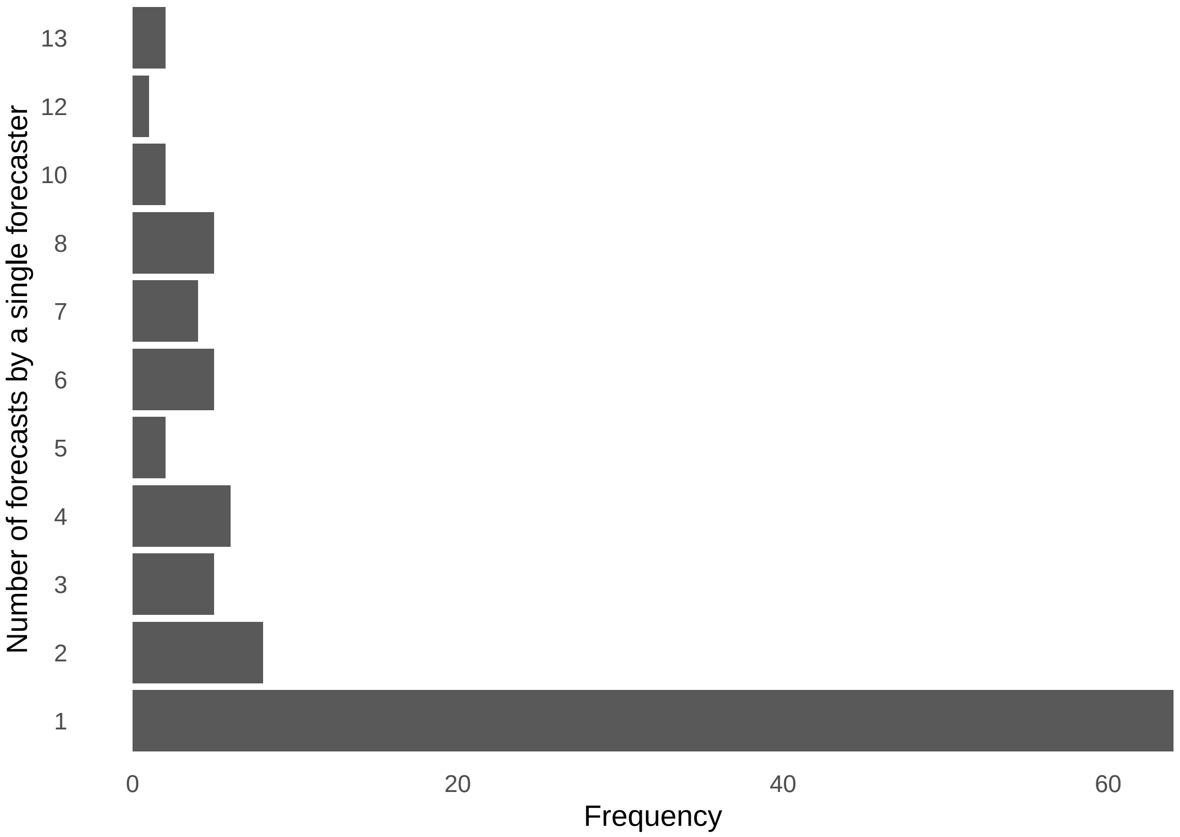
Here are the 10 most active forecasters:

And a summary of the number of forecasts made

And a visualisation of the distribution of the number of forecasts made

model	n__forecasts
anonymous__Stingray	13
seabbs	13
seabbs (Rt)	12
2e10e122	10
BQuilty	10
aurelwu	8
RitwikP	8
seb	8
seb (Rt)	8
Sophia	8

max	min	mean	median
13	1	2.740385	1

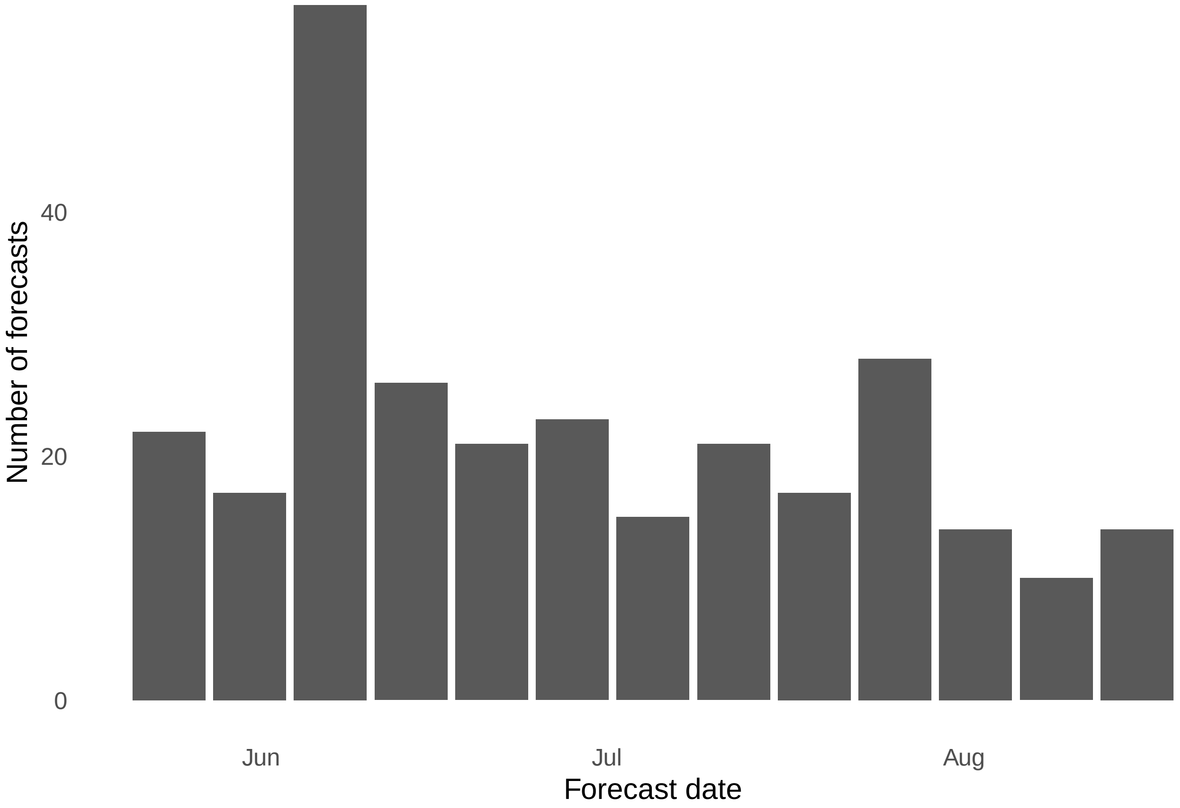


3.2 Number of forecasts per forecast date

Here is a summary of the number of forecasts per forecast date

And the distribution of the number of forecasters over time

max	min	mean	median
57	10	21.92308	21


Number of forecasts
40
20
0
Jun
Jul
Aug
Forecast date