European COVID-19 Forecast Hub: March - August 2021

Kath Sherratt, Sebastian Funk, Johannes Bracher
LSHTM

7 September 2021

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

centre for mathematical modelling of infectious diseases

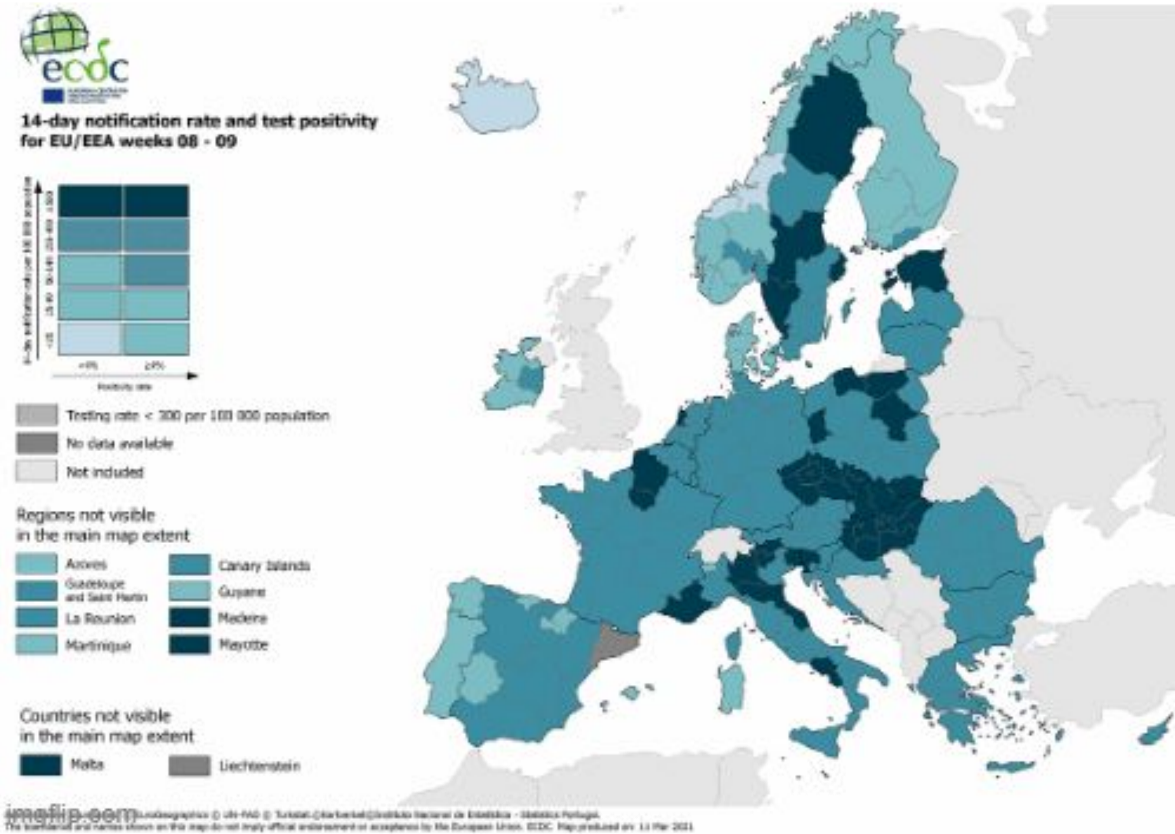
Epidemiological context

cmmid centre for mathematical modelling of infectious diseases

LONDON SCHOOL of HYGIENE & TROPICAL MEDICINE

COVID-19 trends across Europe, 2021

- *March - April*
  - High rates of cases and deaths in Eastern Europe and Sweden
- *May - June*
  - High rates of cases in France
  - Spread of Delta variant in UK
  - Vaccinations beginning to show effec on stabilisting deaths
- *July - August*
  - Spread of Delta variant through Europe
  - High case rates in Spain spreading through France



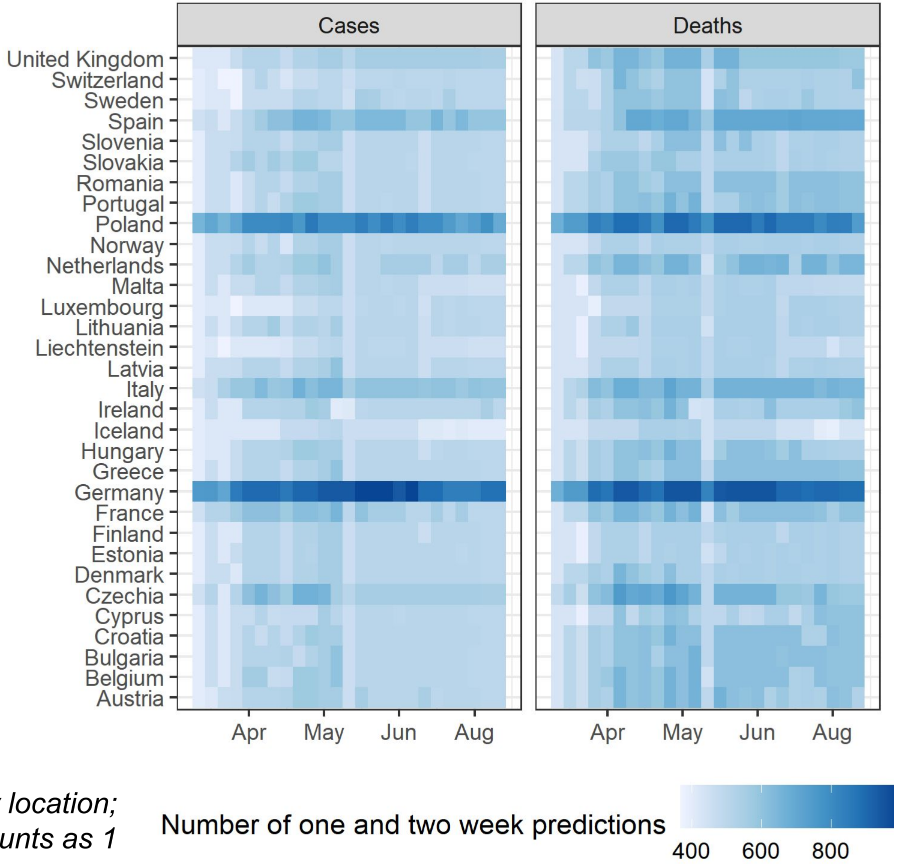
14-day case notification rate per 100,000, and test positivity for EU/EEA Source: ECDC

Hub contributions

How have teams contributed to the Hub?

- Huge volume of contributions
  - **41 models submitted by 34 different teams**
  - 37 models with the full set of predictive quantiles
  - Total of 1,593,444 distinct forecast values submitted between 8 March and 31 August 2021

- **Ensemble** of all forecasts: EuroCOVIDhub-ensemble
  - *8 March - July 2021*: we calculated a **mean** ensemble (each quantile is the mean of all submitted quantiles)
  - *19 July - ongoing:* we switched to a **median** ensemble (each quantile is the median of all submitted quantiles) to be more robust to outlier forecasts
  - We are monitoring the performance of **trained** ensembles that are weighted means/medians



Total number of forecast values each week by location; each quantile of each forecast counts as 1

Comparing forecasts

How can we compare performance between models across multiple parameters?

- Forecast performance = forecasts versus data:
  - Johns Hopkins data
  - Anomalies removed (negative reporting, no data reported)
- Comparisons between models need to account for multiple targets - 2 variables of cases/deaths, of 32 locations, 4 horizons

We used two methods for comparison:

- **Absolute error (point forecasts)**:
  - AE = | observed value - point prediction |
  - Does not consider quantification of uncertainty
- **Weighted interval score (quantile forecasts)**
  - WIS = weighted sum of interval score for each central interval [α, 1-α]

$$\mathrm{IS}_\alpha(F, y) = \underbrace{(u - l)}_{\text{spread}} + \underbrace{\frac{2}{\alpha}(l - y)\mathbb{1}(y < l)}_{\text{penalty for underprediction}} + \underbrace{\frac{2}{\alpha}(y - u)\mathbb{1}(y > u)}_{\text{penalty for overprediction}},$$

  - (see Bracher et al., PLoS Comp Biol 2021, and presentation on evaluating interval forecasts linked at https://covid19forecasthub.eu/community.html)
  - Penalises wide forecasts as well as ones that are far from the data

Systematic comparison

cmmid centre for mathematical modelling of infectious diseases

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

- Models are assessed relative to a **baseline** forecast

1. Relative “skill” (via mean WIS/AE) is computed between **each pair of models**

$$\theta_{ij} = \frac{\text{mean WIS model } i \text{ on } \mathcal{A}_{ij}}{\text{mean WIS model } j \text{ on } \mathcal{A}_{ij}}$$

with $\mathcal{A}_{ij}$ as the overlap of available forecasts by i and j and

2. Each model has a relative skill as the geometric mean of **all pairwise skills**

$$\theta_i = \left(\prod_{m=1}^{M} \theta_{im} \right)^{1/M}$$

3. A re-scaled relative skill is obtained by comparing to a **baseline model**

$$\theta_i^* = \frac{\theta_i}{\theta_B},$$

where θ_B is the relative WIS skill of the baseline model.

Approach developed by Bracher and others for the US Forecast hub; see Cramer et al. (2021)

Evaluation Germany

CSV Excel

model	n	rel_wis	rel_ae
itwm-dSEIR	26	0.52	0.51
EuroCOVIDhub-ensemble	26	0.54	0.59
MUNI-ARIMA	18	0.56	0.6
HZI-AgeExtendedSEIR	25	0.57	0.74
epiforecasts-EpiExpert_direct	19	0.67	0.67
ILM-EKF	26	0.69	0.81
epiforecasts-EpiExpert	26	0.7	0.77
Karlen-pypm	26	0.79	0.83
LANL-GrowthRate	25	0.81	0.7
UNIPV-BayesINGARCHX	25	0.83	0.6[illegible]

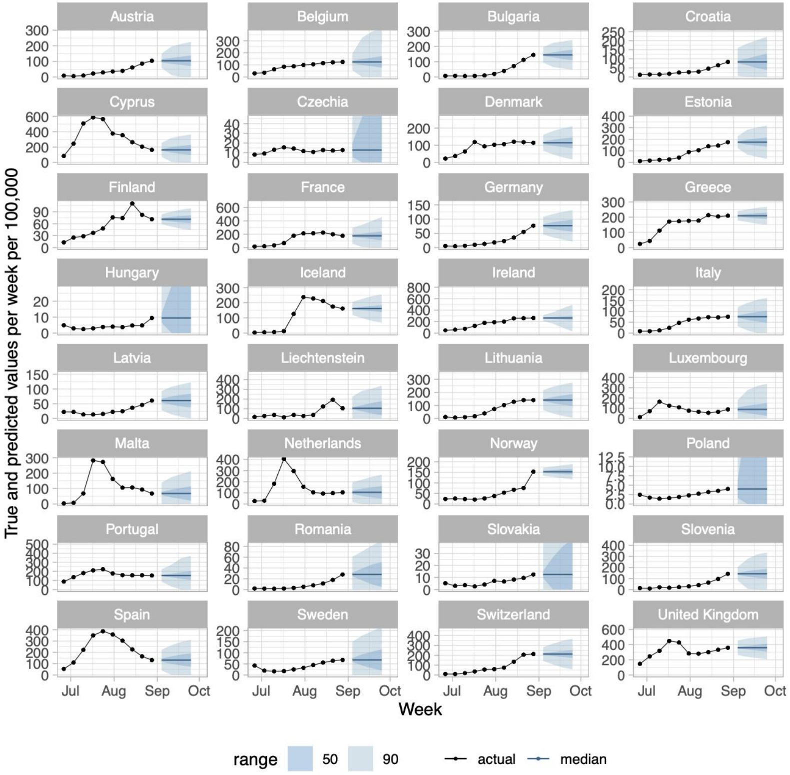
http://covid19forecasthub.eu/reports.html

Relative skill: interpretation

- Interpretation: a model is better than the baseline model if its **relative skill is <1**.

- Note: this is not the same as a direct comparison to the baseline as it **accounts for how difficult it is to beat the baseline** on the targets that the model addressed

- Baseline forecast: "**same incidence next week as this week**"
    - Expanding uncertainty over time, informed by past differences in incidence
    - Developed and used by the US COVID-19 forecast hub (Cramer et al., 2021).



Baseline model forecasts of 31 August 2021.

Forecast performance

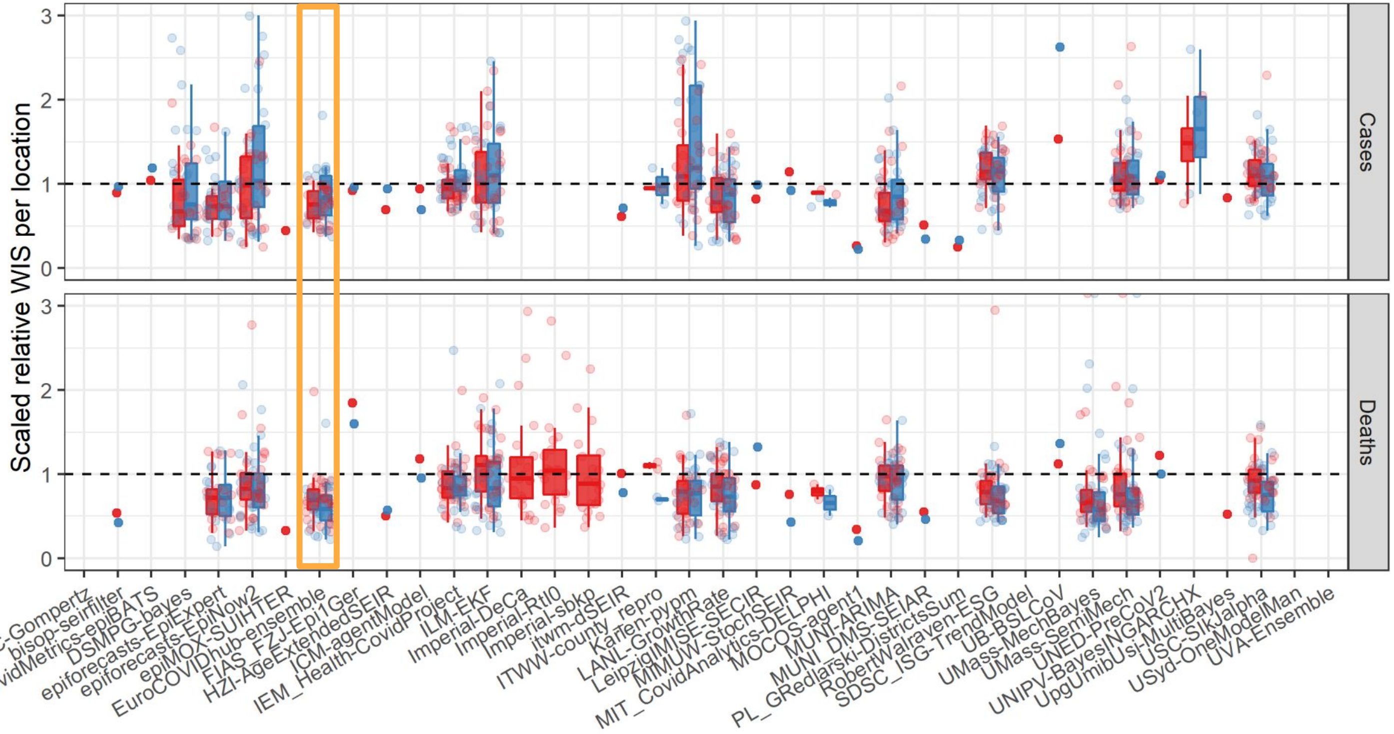
Relative performance: WIS

centre for mathematical modelling of infectious diseases

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

How do forecasts perform relative to the baseline? Comparison of relative weighted interval score

- WIS only calculated for models with full range of quantiles (34)
- Better performance is relative to the baseline: <1
- Better performance and less variance when forecasting **deaths**, compared to cases
- Similar performance **across horizons** (slightly better average performance at 1 week than 2)
- **Ensemble** consistently outperforms baseline for both cases and deaths



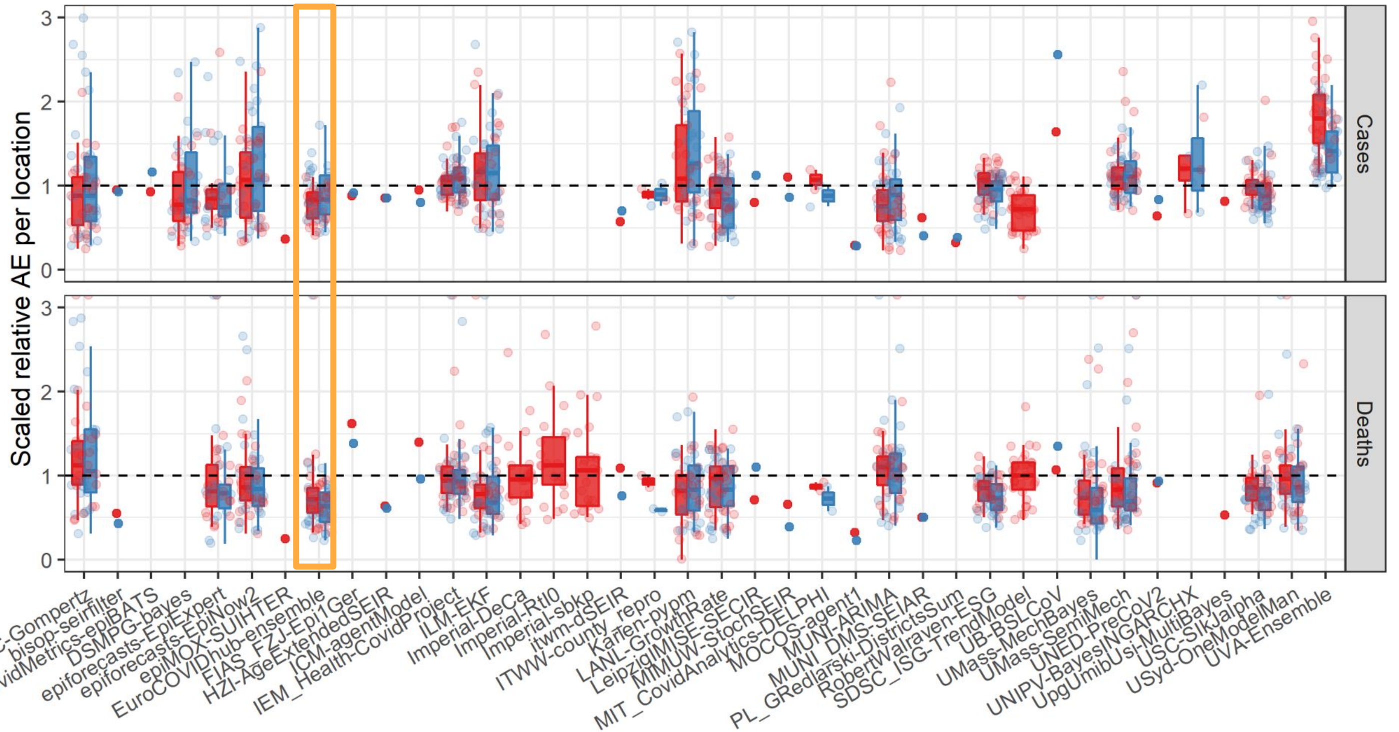
34 models' relative weighted interval score; points represent score for each location, with boxplot for distribution across multiple locations (plot limited to scores <3). Ensemble highlighted in yellow.

Weeks ahead 1 2

Relative performance: AE

How do forecasts perform relative to the baseline? Comparison of relative absolute error

- Calculated on median/point prediction (all 40 models included)
- Better performance is relative to the baseline, <1
- Strongly correlated to WIS for models with uncertainty
- **Ensemble** still beats baseline; appears slightly less consistent across locations



All models' relative absolute error; points represent score for each location, with boxplot for distribution across multiple locations (plot limited to scores <3). Ensemble highlighted in yellow.

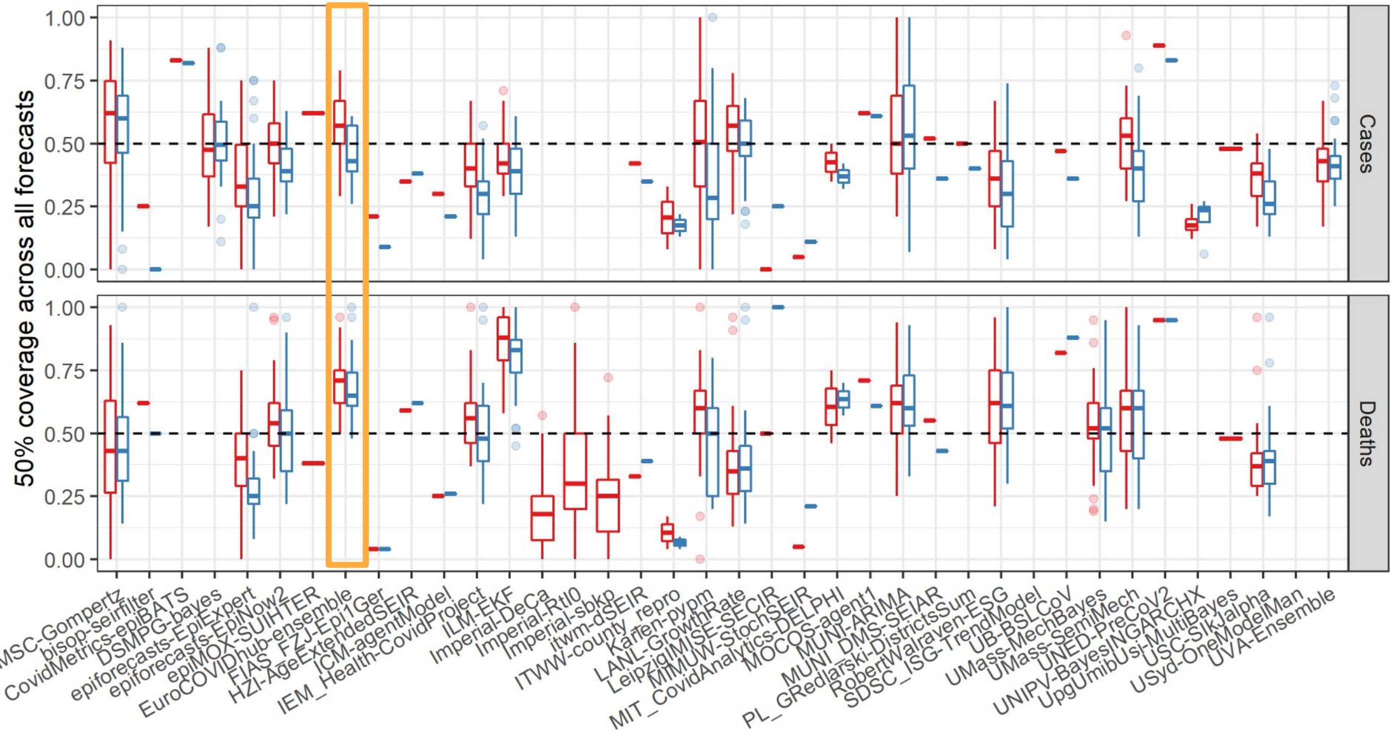
Coverage of uncertainty

cmmid centre *for* mathematical modelling *of* infectious diseases

LONDON SCHOOL *of* HYGIENE &TROPICAL MEDICINE

How accurately calibrated are probabilistic predictions?

- Most models (39, 95%) included some uncertainty
- A perfect forecast would achieve 50% coverage of observations at the 0.50 prediction interval
- Coverage slightly more accurate for cases: average coverage **20-89%**
- Uncertainty for deaths fell across near the entire spectrum: **4-95%**
- **Ensemble** relatively underconfident:
    - 57% for cases
    - 71% for deaths



The proportion of observations that fell within the 50% prediction interval for each model, by target count of cases and deaths and horizon.

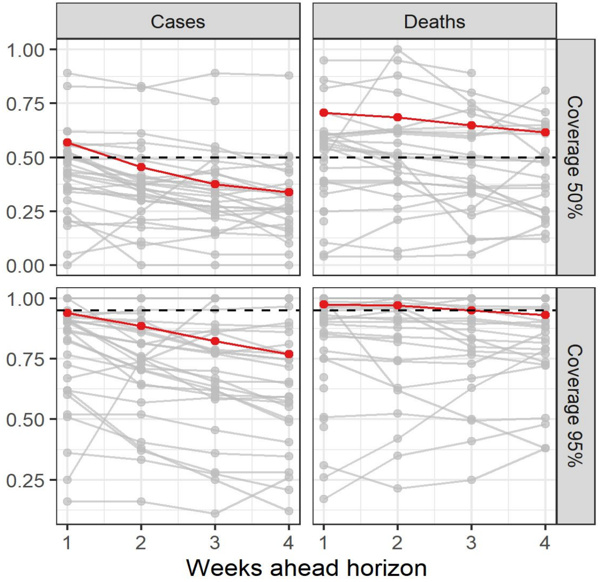
Weeks ahead 1 2

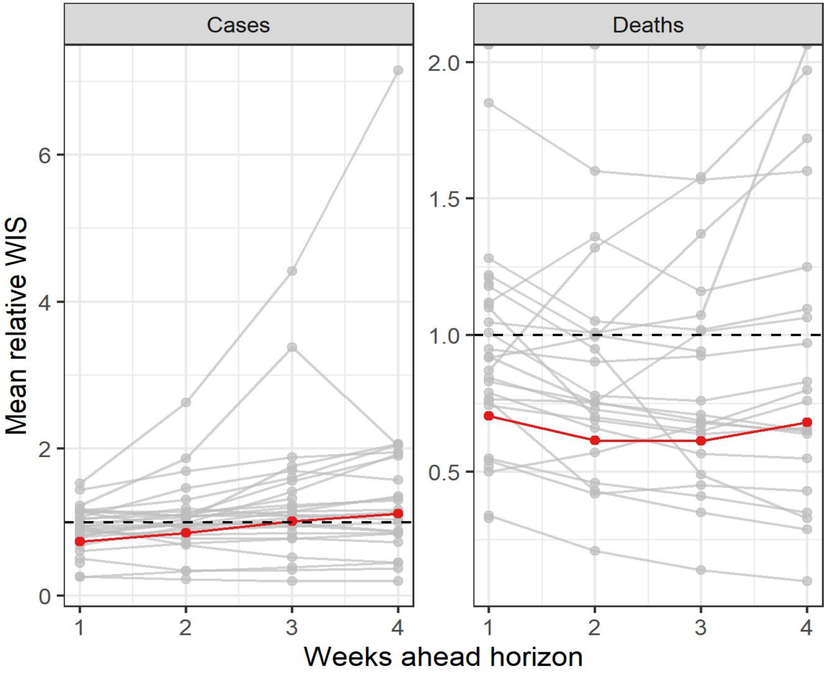
Forecasting over horizons

How does performance change further into the future?

- **Coverage** worsened slightly at longer horizons (averaging 41% and 51% for two-week case and death forecasts respectively).
- **Relative WIS** worsened at 3-4 weeks for cases
- **Ensemble** still outperformed baseline for deaths

50% and 95% coverage of each model across all locations by horizon, relative to ideal coverage of 0.5 and 0.95; ensemble forecast in red



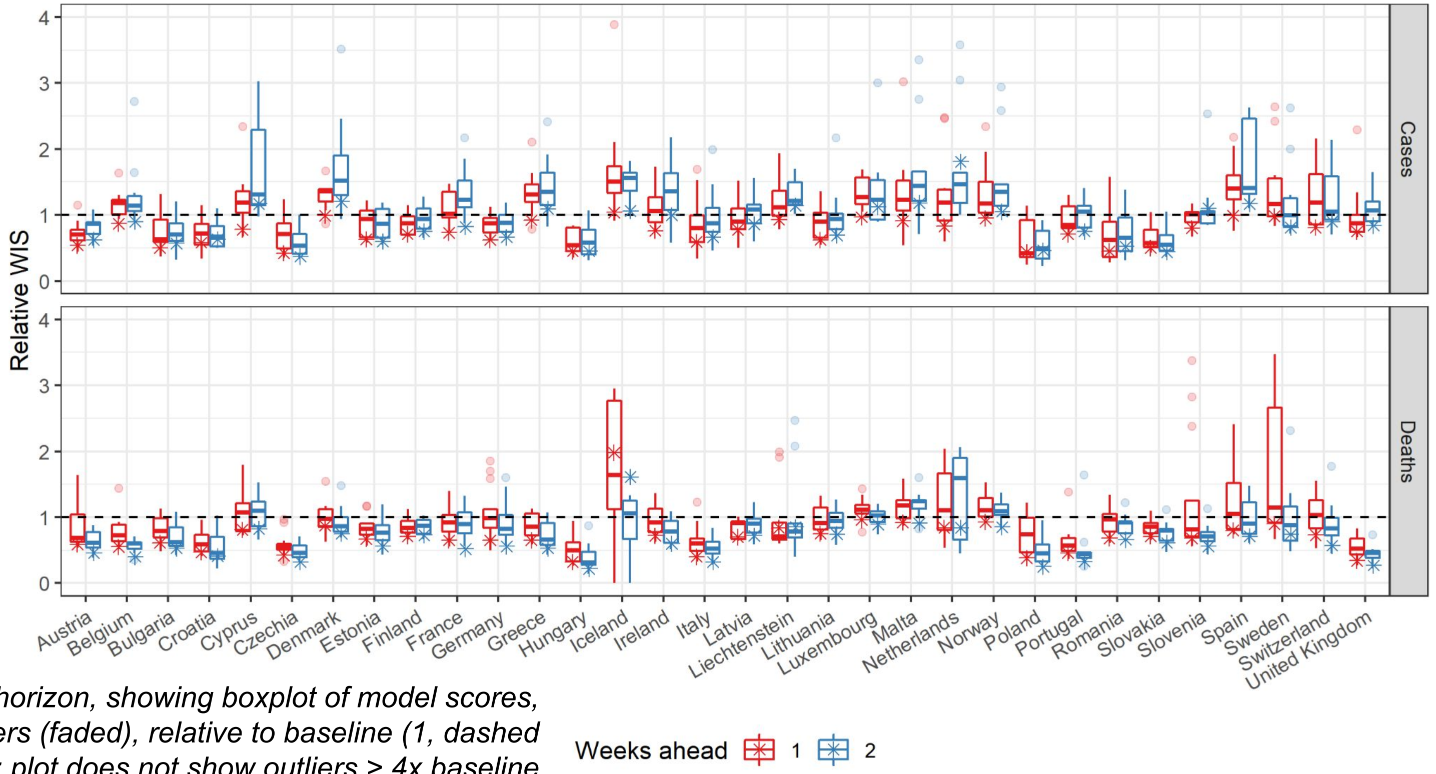


Relative WIS for each model across all forecast locations by horizon, relative to baseline forecast; ensemble forecast in red

Forecasting by country

Are some countries easier or harder for models to predict than others?

- Better performance of models relative to baseline is <1
- Average scores by country were roughly equivalent to baseline score
- Countries with **very low absolute counts** had wider errors compared to baseline
  - Cyprus, Iceland, Netherlands
- **Ensemble** (asterisk) generally among the best models in each country



Relative WIS by country and horizon, showing boxplot of model scores, ensemble (asterisk), and outliers (faded), relative to baseline (1, dashed line); plot does not show outliers > 4x baseline

Next steps

Future work

centre for mathematical modelling of infectious diseases

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

- **Hospitalisations**
  - So far only a few teams
  - More contributions welcome
  - We expect this to become the most important target to ECDC and national health agencies

- **Trained ensembles**
  - Ongoing work
  - Conclusion from other hubs: unweighted median difficult to beat

- **Community**
  - Exploring ways to give more individual feedback to teams

Summary

- **Performance highlights**
  - Models out-performed the baseline at **short (1-2 week) horizons** and for **death forecast targets**
  - The **ensemble** of all models is the most reliably well-performing model across locations

- We are writing these results into a **manuscript** to be shared with all teams for comments

- We welcome your **independent analysis** of forecasts:
  - All data, code, **downloadable** from Github
  - We use R packages `covidHubUtils` to navigate around forecasts and observed data, and `scoringutils` to evaluate forecasts

Thanks to collaborators:

- ECDC team: Helen Johnson, Rene Niehus, Rok Grah
- Johannes Bracher and team at Karlsruhe Institute of Technology (KIT)
- Nick Reich, Evan Ray and the US Forecast Hub team at University of Massachusetts (UMass) Amherst
- Signale team at the Robert-Koch Institute