# Using human insight and computational methods to improve real-time epidemic forecasting

- PhD Upgrading Seminar -

Nikos Bosse

22nd July 2021

# Supervisors

Sebastian Funk (LSHTM)

Anne Cori (Imperial)

Edwin van Leuuwen (PHE)

# Advisory Committee

Sam Abbott (LSHTM)

Johannes Bracher (KIT)

# Outline

**Background, aim and objectives**

**Paper I** - Evaluating epidemiological forecasts

**Paper II** - Comparing human and model-based forecasts

**Paper III** - A deeper understanding of human predictions

**Paper IV** - Optimal ensembles in epidemiological forecasting

**Timetable**

# Background, aim and objectives

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

# PhD - Background

- Forecasting epidemics is important for public health policy
- Improving forecasts requires evaluating and understanding past forecasts
- Forecasts are usually a mix of human insight and model-based assumptions
- Collaborative efforts play an important role

  → Forecast Hubs in US, Germany & Poland, Europe

  → individual predictions need to be aggregated

# PhD - Aim

**Aim**: to improve infectious disease forecasting by obtaining a better understanding of

- the way in which forecasts can best be evaluated
- the role of human insight in infectious disease forecasting and how human predictions and model-based approaches can best be combined
- how the optimal choice for an ensemble method depends on the characteristics and the number of available models

# PhD - Objectives

- Establish appropriate tools to evaluate predictions and summarise best practices in forecast evaluation
- Collect human predictions of COVID-19 in Germany, Poland and the UK
- Analyse characteristics of human predictions and compare them against model-based forecasts to discern relative strengths and weaknesses
- Create and evaluate a hybrid forecasting approach which combines human insight and model-based inference
- Create ensembles of differing sizes to identify characteristics of optimal ensembles

# Paper I

Evaluating epidemiological forecasts - tools and best practices

- Background and motivation
- Aim and objectives
- Project outline
- Overview: forecast evaluation
- Current progress

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

# Paper I - Background and motivation

- Forecast evaluation is necessary to improve forecast methodology
- There is an extensive literature on possible metrics (e.g. Gneiting and Raftery 2007), but
  - no easy-to-use tools
  - little discussion on which metrics are appropriate in which context (esp. in epidemiological forecasting)

# Paper I - Aim and objectives

**Aim**: to improve and facilitate the evaluation of forecasts

**Objectives**:

- to create tools that facilitate evaluating forecasts using the statistical software *R*
- to review and discuss existing methods and how they can best be applied in an epidemiological context

# Paper I - Introduction

**Point forecast**:

"Tomorrow at 10am, it will be exactly 20.4°C" → no uncertainty

**Probabilistic forecast**:

"Tomorrow at 10am it will be between 18°C and 22°C" → expresses uncertainty

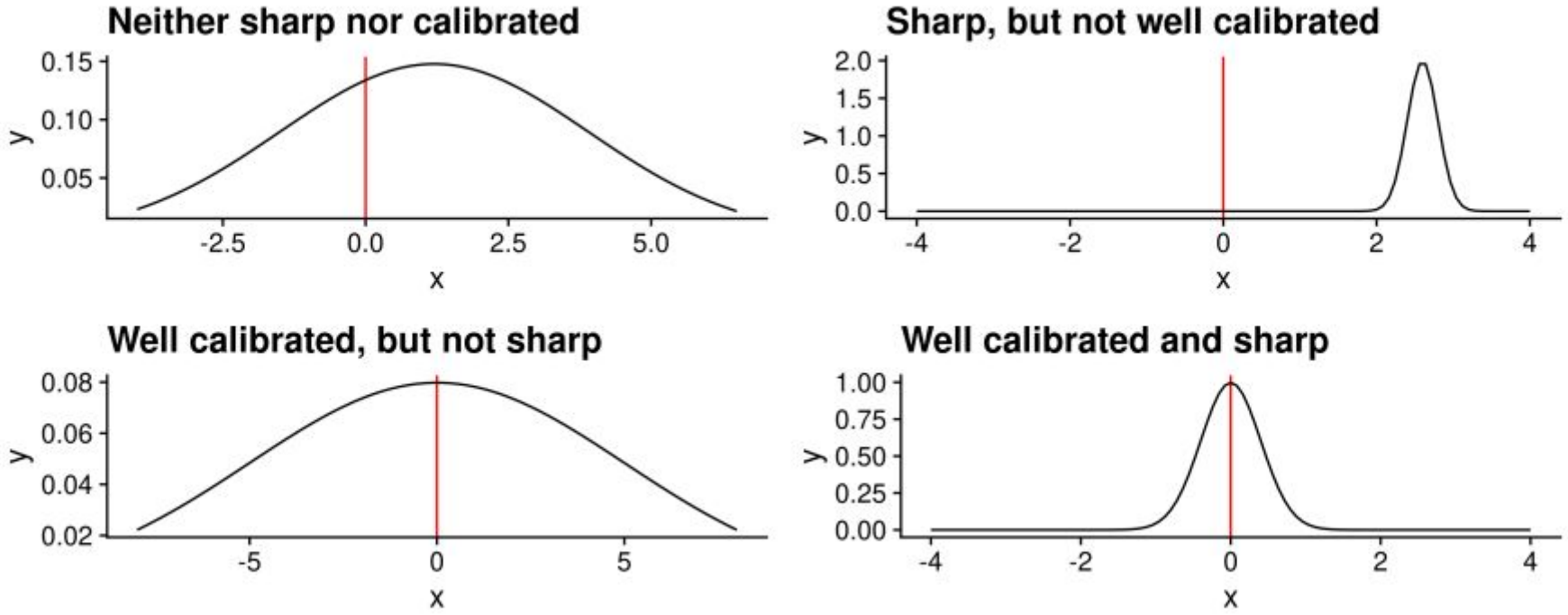
"Between 20.3°C and 20.4°C" → probably wrong, but useful if true

"Between 10°C and 30°C" → probably right, but not very useful

# Paper I - Introduction

**Forecasting paradigm** (Gneiting et al. 2005; Gneiting, Balabdaoui, and Raftery 2007)

"Maximise sharpness of the predictive distribution subject to calibration"



# Paper I - Introduction

**Assessing sharpness and calibration independently:**

- Helpful for model diagnostics

**Proper scoring rules**:

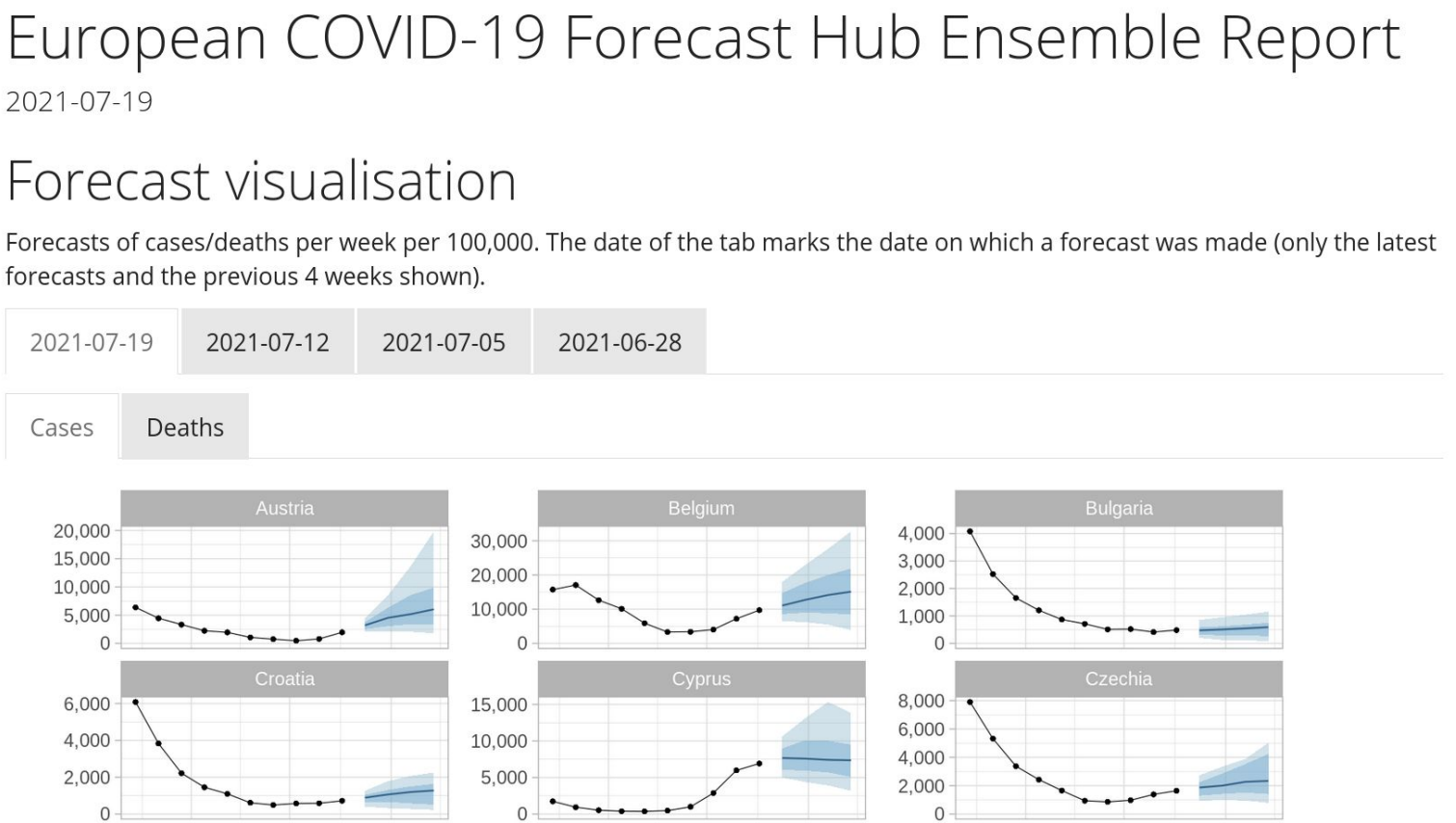
- Summarise the trade-off between sharpness and calibration in a single number
- The 'correct' model always scores best → The score can't be cheated
- Helpful for creating a ranking between forecasters
- One example: weighted interval score (WIS) (Bracher et al. 2021)

# Paper I - Methods

- The *R* package *scoringutils* implements
  - proper scoring rules
  - metrics that allow to analyse sharpness and calibration in detail
  - pairwise comparisons, to compare models even if there are missing observations
  - generic plotting functions to visualise the output
- The accompanying paper
  - summarises the existing literature on forecast evaluation and discusses when and how to use which metric
  - evaluates short-term predictions of COVID-19 in the UK

# Paper I - Results

- The *scoringutils* package is currently used to create visualisations and evaluations for the European Forecast Hub (and scores for the US Forecast Hub)



# Paper I - Results

- The *scoringutils* package is currently used to create visualisations and evaluations for the European Forecast Hub (and scores for the US Forecast Hub)

## Forecast scores

Scores separated by target and forecast horizon. Only models with submissions in each of the last 4 weeks are shown.

Cases | Deaths

1 week ahead horizon | 2 weeks ahead horizon | 3 weeks ahead horizon | 4 weeks ahead horizon

CSV | Excel

Search:

| | model | n | rel_skill | 50% Cov. | 95% Cov. | wis | sharpness | underpred | overpred | bias | aem |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | IEM_Health-CovidProject | 19 | 0.35 | 0.42 | 0.79 | 4807.49 | 1335.31 | 2797.65 | 674.53 | -0.3 | 7074.74 |
| 2 | epiforecasts-EpiExpert_Rt | 12 | 0.39 | 0.58 | 1 | 6227.14 | 4251.86 | 1060.7 | 914.57 | -0.3 | 9632.42 |
| 3 | epiforecasts-EpiExpert | 19 | 0.39 | 0.32 | 0.79 | 5589.31 | 2202.37 | 1466.14 | 1920.81 | -0.02 | 8785.37 |
| 4 | LANL-GrowthRate | 18 | 0.4 | 0.28 | 0.67 | 5562.97 | 1236.87 | 3832.15 | 493.95 | -0.36 | 7282.78 |
| 5 | EuroCOVIDhub-ensemble | 19 | 0.41 | 0.42 | 0.95 | 5499.53 | 2192.66 | 2650.68 | 656.19 | -0.29 | 8653.32 |

# Paper I - Results

- The paper has a detailed explanation of every metric

| Metric | Explanation |
|---|---|
| CRPS (Continuous) ranked probability score | The crps is a proper scoring rule that generalises the absolute error to probabilistic forecasts. It measures the 'distance' of the predictive distribution to the observed data-generating distribution. The CRPS is given as $$\text{CRPS}(F, y) = \int_{-\infty}^{\infty} (F(x) - 1(x \geq y))^2 \, dx,$$ where y is the true observed value and F the CDF of predictive distribution. Often An alternative representation is used: $$\text{CRPS}(F, y) = \frac{1}{2}\mathbb{E}_F\lvert X - X'\rvert - \mathbb{E}_F\lvert X - y\rvert,$$ where $X$ and $X'$ are independent realisations from the predictive distributions $F$ with finite first moment and $y$ is the true value. In this represenation we can simply replace $X$ and $X'$ by samples sum over all possible combinations to obtain the CRPS. For integer-valued forecasts, the RPS is given as $$\text{RPS}(F, y) = \sum_{x=0}^{\infty} (F(x) - 1(x \geq y))^2.$$ **Usage and caveats** Smaller values are better. The crps is a good choice for most practical purposes that involve decision making, as it takes the entire predictive distribution into account. If two forecasters assign the same probability to the true event $y$, then the forecaster who assigned high probability to events far away from $y$ will still get a worse score. The crps (in contrast to the log score) can at times be quite lenient towards extreme mispredictions. Also, due to it's similarity to the absolute error, the level of scores depend a lot on the absolute value of what is predicted, which makes it hard to compare scores of forecasts for quantities that are orders of magnitude apart. |

*(continued)*

| Metric | Explanation |
|---|---|
| Log score | The Log score is a proper scoring rule that is simply computated as the log of the predictive density evaluated at the true observed value. It is given as $$\text{log score} = \log f(y),$$ where $f$ is the predictive density function and y is the true value. For integer-valued forecasts, the log score can be computed as $$\text{log score} = \log p_y,$$ where $p_y$ is the probability assigned to outcome p by the forecast F. **Usage and caveats**: Larger values are better, but sometimes the sign is reversed. The log score is ensitive to outliers, as individual negative log score contributions quickly can become very large if the event falls in the tails of the predictive distribution, where $f(y)$ (or $p_y$) is close to zero. Whether or not that is desirable depends ont the application. In scoringutils, the log score cannot be used for integer-valued forecasts, as the implementation requires a predictive density. In contrast to the crps, the log score is a local scoring rule: it's value only depends only on the probability that was assigned to the actual outcome. This property is desirable for inferential purposes, for example in a Bayesian context (@winklerScoringRulesEvaluation1996). In settings where forecasts inform decision making, it may be more appropriate to score forecasts based on the entire predictive distribution. |

*(continued)*

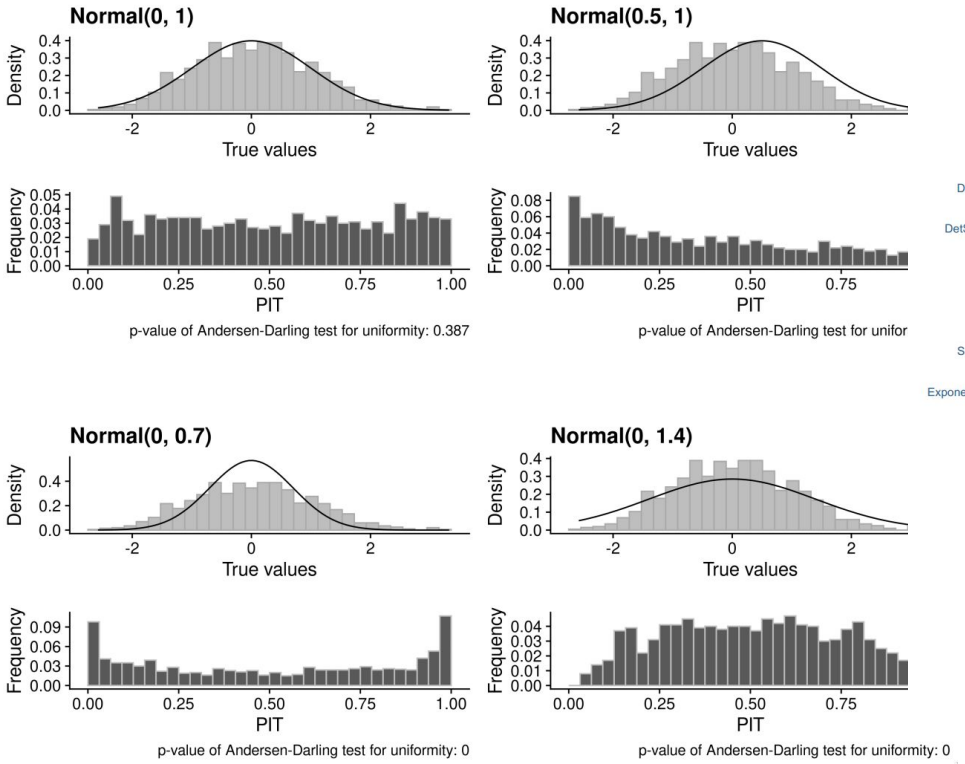
| Metric | Explanation |
|---|---|
| WIS (Weighted) interval score | The (weighted) interval score is a proper scoring rule for quantile forecasts that converges to the crps for an increasing number of intervals. The score can be decomposed into a sharpness (uncertainty) component and penalties for over- and underprediction. For a single interval, the score is computed as $$IS_\alpha(F, y) = (u - l) + \frac{2}{\alpha} \cdot (l - y) \cdot 1(y \leq l) + \frac{2}{\alpha} \cdot (y - u) \cdot 1(y \geq u),$$ where 1() is the indicator function, $y$ is the true value, and $l$ and $u$ are the $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$ quantiles of the predictive distribution $F$, i.e. the lower and upper bound of a single prediction interval. For a set of $K$ prediction intervals and the median $m$, the score is computed as a weighted sum, $$WIS = \frac{1}{K + 0.5} \cdot (w_0 \cdot \lvert y - m\rvert + \sum_{k=1}^{K} w_k \cdot IS_\alpha(F, y)),$$ where $w_k$ is a weight for every interval. Usually, $w_k = \frac{\alpha_k}{2}$ and $w_0 = 0.5$. **Usage and caveats**: Smaller scores are better. Applicable to all quantile forecasts, takes the entire predictive distribution into account. Just as the crps, the wis is based on measures of absolute error. When averaging across multiple targets, it will therefore be dominated by targets with higher absolute values. The decomposition into sharpness, over- and underprediction make it easy to interpret scores and use them for model improvement. |

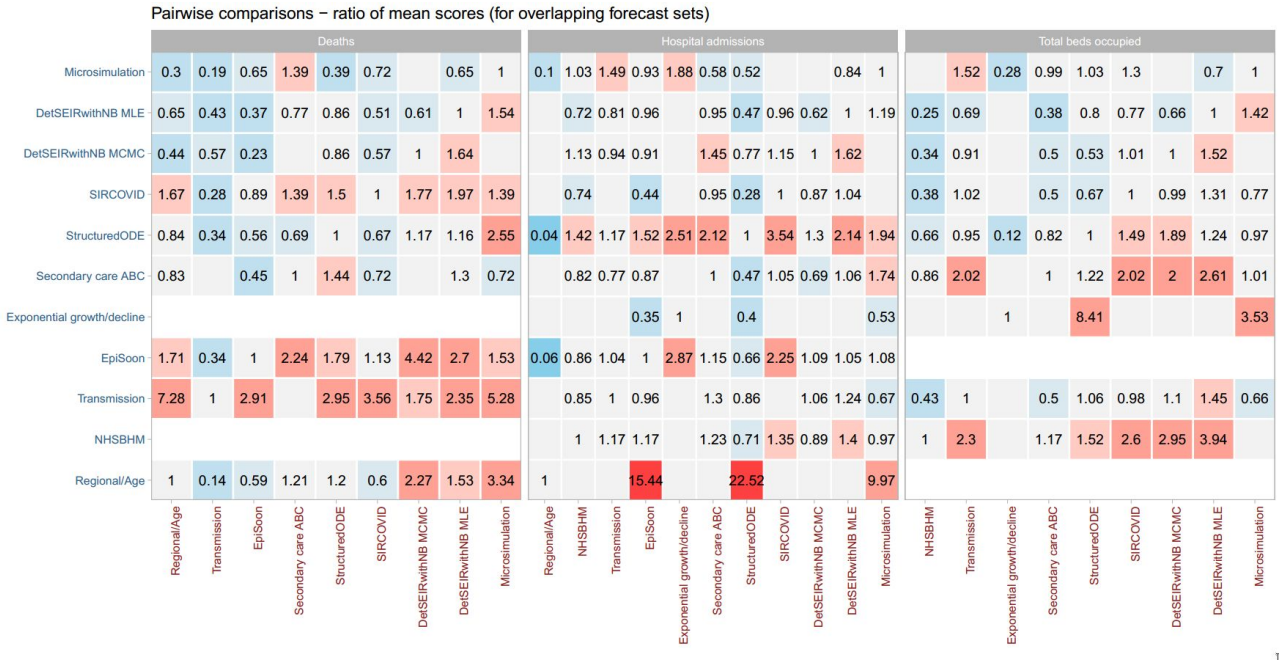
*(continued)*

| Metric | Explanation |
|---|---|
| DSS Dawid-Sebastiani score | The Dawid-Sebastiani-Score is a proper scoring rule proposed by Gneiting and Raftery in [@gneitingStrictlyProperScoring2007] that only relies on the first moments of the predictive distribution and is therefore easy to compute. It is given as $$\text{dss}(F, y) = \left(\frac{y - \mu}{\sigma}\right)^2 + 2 \cdot \log \sigma,$$ where $F$ is the predictive distribution with mean $\mu$ and standard deviation $\sigma$ and $y$ is the true observed value. textbfUsage and caveats The dss is applicable to continuous and integer forecasts and easy to compute. Apart from the ease of computation we see little advantage in using it. MAYBE SCRATCH IT ALTOGETHER FROM THE PAPER? |
| Brier score | Proper scoring rule for binary forecasts. The Brier score is computed as $$\text{Brier Score} = \frac{1}{N} \sum_{n=1}^{N} (f_n - y_n),$$ where $f_n$, with $n = 1, \ldots, N$ are the predicted probablities that the corresponding events, $y_n \in (0, 1)$ will be equal to one.) **Usage**: Applicable to all binary forecasts. |

# Paper I - Results

- It showcases visualisations and discusses their interpretation



# Paper I - Current progress

- Early draft of the paper is written
- R package is operational (github.com/epiforecasts/scoringutils), but needs some more work pre-publication

# Paper II

Evaluating human predictions of Covid-19 against epidemiological model forecasts in Germany and Poland

- Background and motivation
- Aim and objectives
- Methods
- Results
- Current progress

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

# Paper II - Background and motivation

- Large amounts of time and resources go into developing forecasting models
- Unclear whether models add anything beyond the obvious
- Many crowd forecasting efforts exist
- BUT: no direct comparison to model-based predictions

# Paper II - Aim and objectives

**Aim**: to obtain a better understanding of the strengths and weaknesses of human predictions and model-based approaches

**Objectives**:

- to create tools necessary to elicit human time series predictions
- to collect human predictions of cases and deaths from COVID-19 in Germany and Poland
- to compare an ensemble of human opinion with *untuned* model-based forecasts

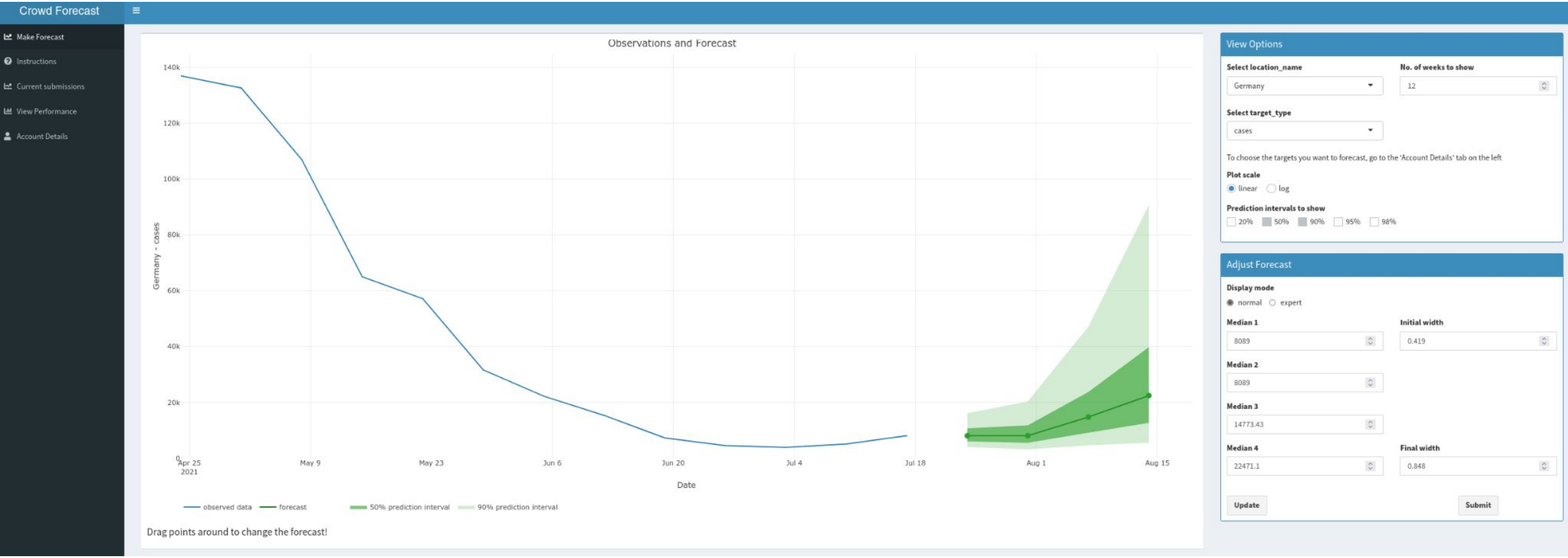
# Paper II - Methods

- Submitted forecasts to the German and Polish Forecast Hub (10/2020 - 03/2021):
  - Human predictions, using web app I developed
  - 2 completely untuned model-based approaches
    - Renewal model

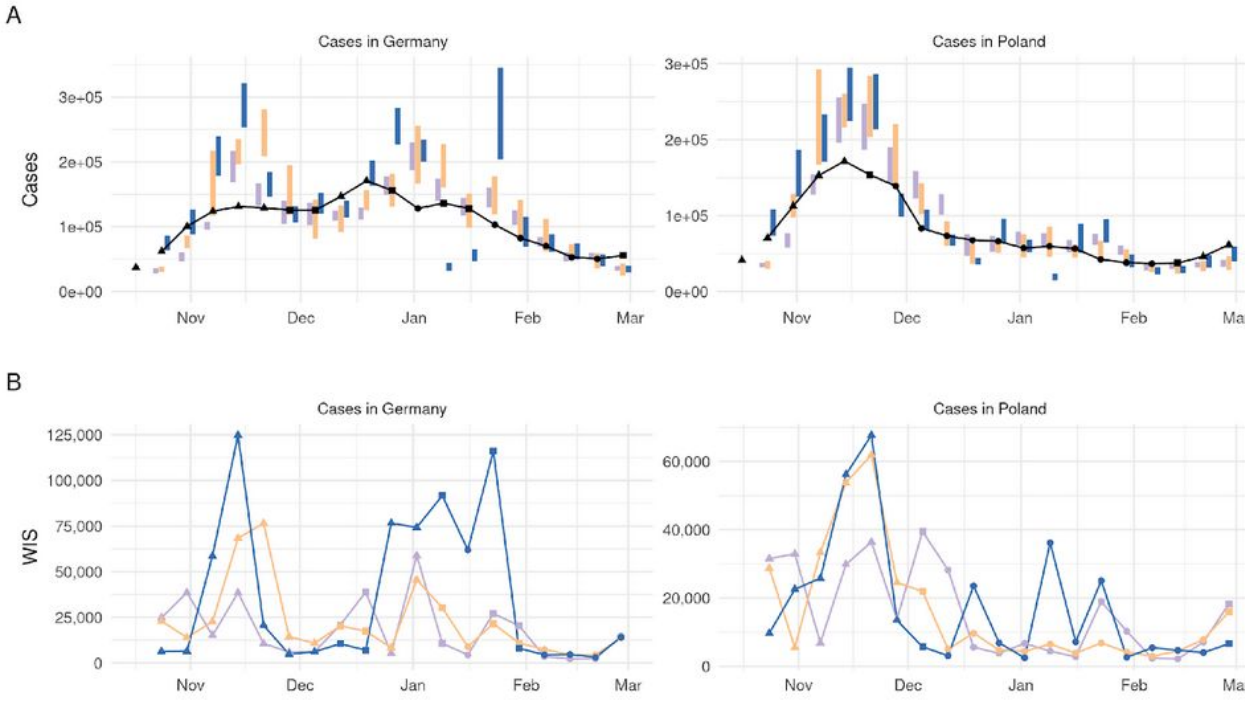
      (future cases = effective reproduction number * cases today) (Abbott et al. 2020)
    - Convolution model

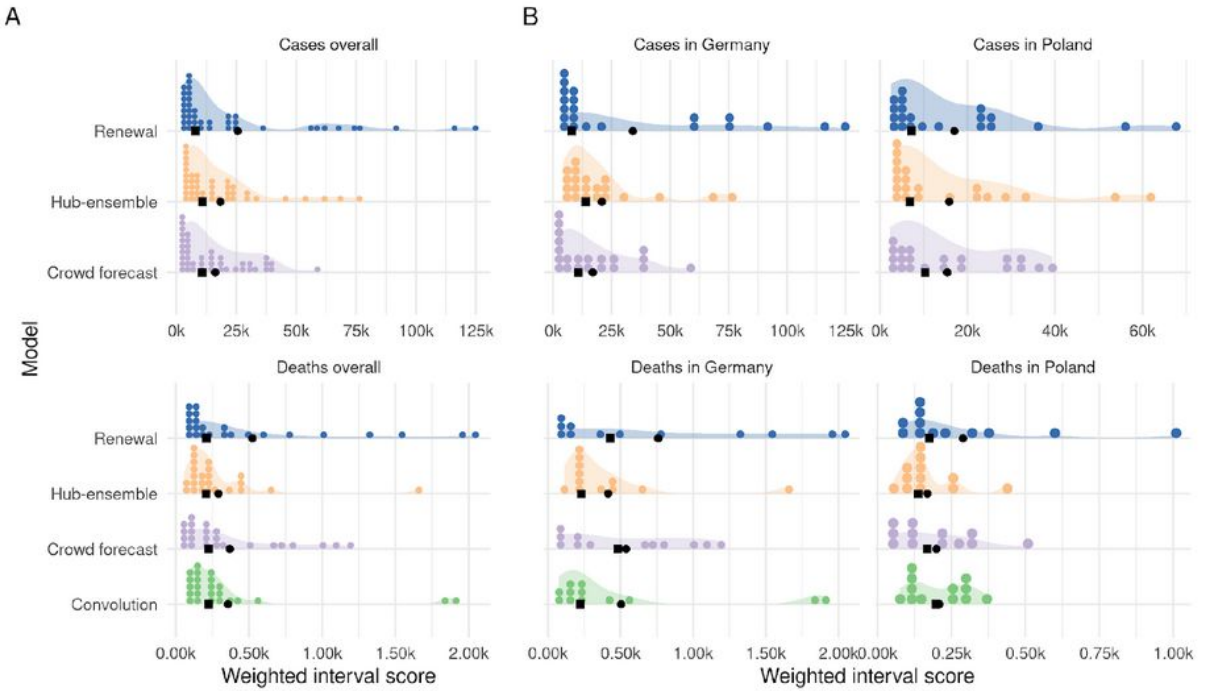
      (future deaths = convolution of past cases)

# Paper II - Methods



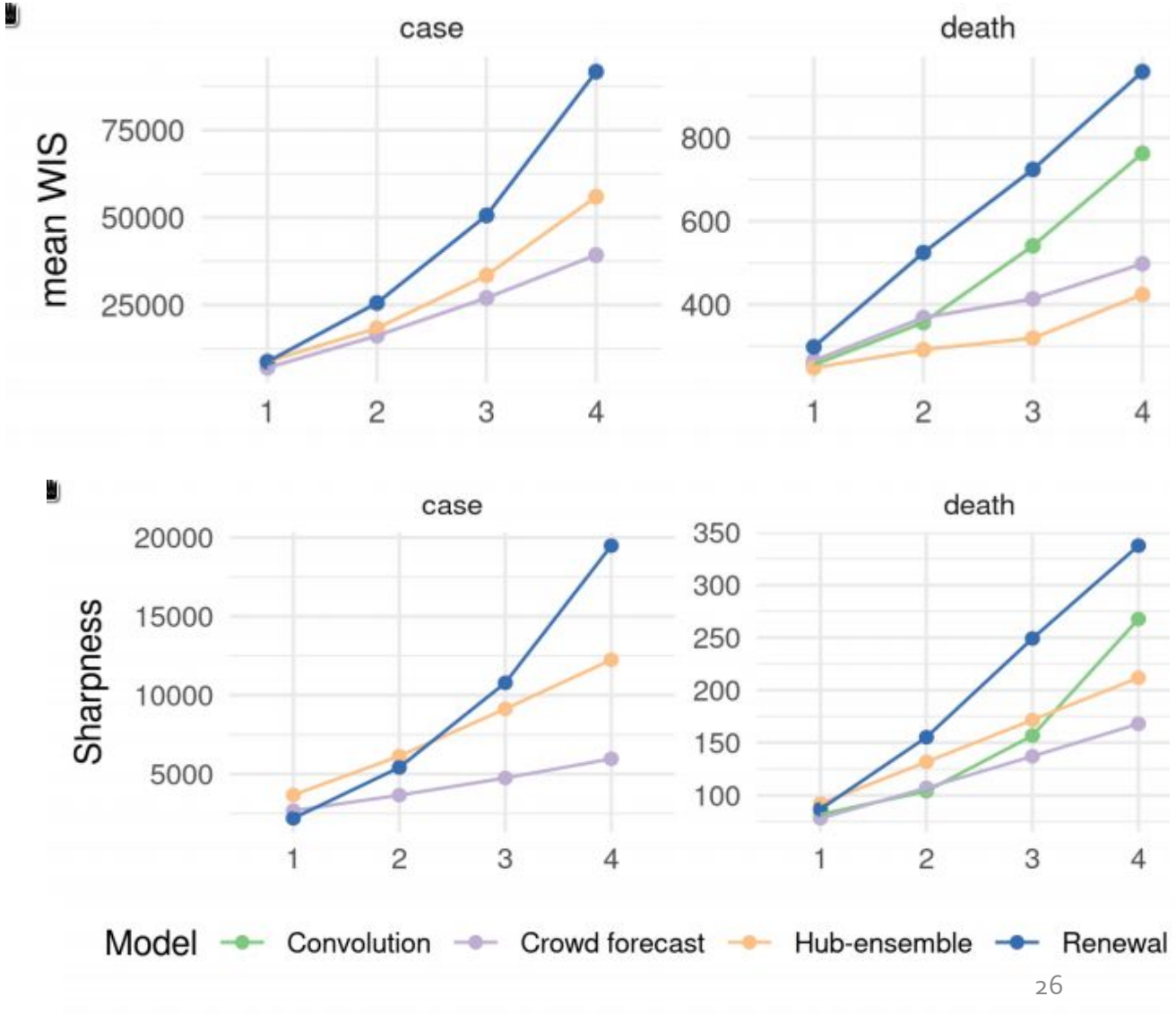
# Paper I - Results




# Paper I - Results

- Humans did relatively well for cases, but less well for deaths
- Humans were overconfident
- Untuned model-based approaches performed well short-term
- Average performance was strongly influenced by outlier predictions



# Paper II - Discussion

- Limited number of participants
- Focus only on an ensemble of human predictions, not on individual forecasters
- We can’t know how much other models submitted to the Forecast Hub were informed by human opinion

# Paper II - Current progress

- Ethics approval received
- Data collected
- Data analysis 80% done
- First draft written

# Paper III

Human predictions of cases, deaths and $R_t$ during the third wave (May - August 2021) of COVID-19 in the United Kingdom

- Background and motivation
- Aim and objectives
- Methods
- Preliminary results
- Discussion
- Current progress

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

# Paper III - Background and motivation

- Previous study showed promising results, but was limited
  - due to the small number of participants
  - only the mean ensemble of human forecasters was analysed

→ Replication with more participants in a different setting

→ Analyse individual forecasters

→ Analyse ways in which models and humans can be combined

# Paper III - Aim and objectives

**Aim**: to obtain a better understanding of human forecasters, and to explore ways in which their insight can be augmented by modelling

**Objectives**:

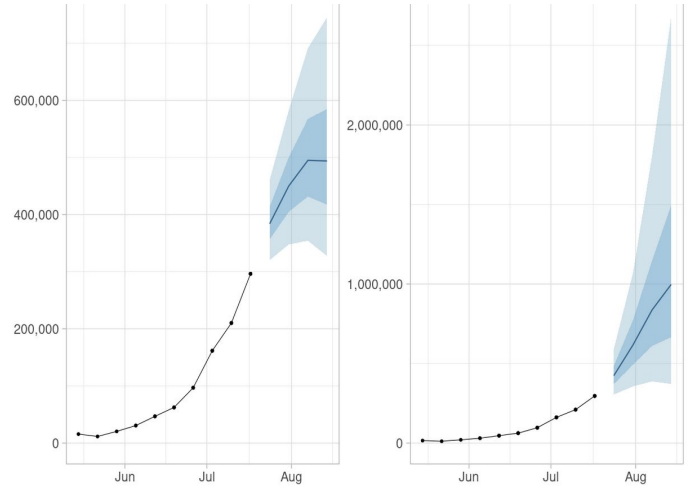
- to collect human predictions of reported cases, deaths, and reproduction numbers
- to analyse individual predictions
- to compare direct forecasts against a hybrid forecasting approach

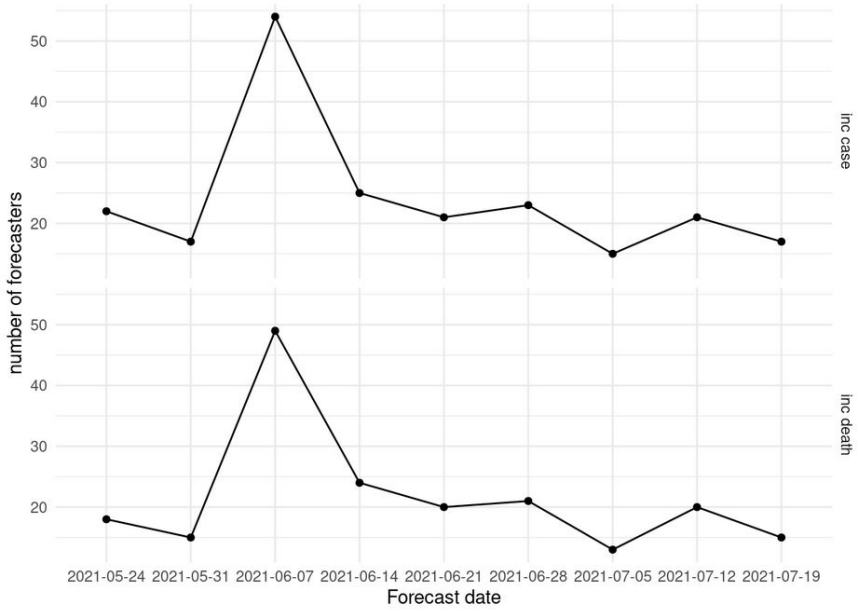
# Paper III - Methods

- Data collected over 12 weeks from May 24th to August 16th 2021
- We collected
  - Direct forecasts (cases, deaths)
  - Forecasts of the effective reproduction number (which were then mapped to cases and deaths using the renewal equation

# Paper III - Preliminary results

- Currently around 20 forecasters, fewer for Rt
- Large heterogeneity in individual forecasts
- Rt forecasts and direct forecasts often differ noticeably




| | model | n | rel_skill | 50% Cov. | 95% Cov. | wis | sharpness | underpred | overpred | bias | aem |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | IEM_Health-CovidProject | 19 | 0.35 | 0.42 | 0.79 | 4807.49 | 1335.31 | 2797.65 | 674.53 | -0.3 | 7074.74 |
| 2 | epiforecasts-EpiExpert_Rt | 12 | 0.39 | 0.58 | 1 | 6227.14 | 4251.86 | 1060.7 | 914.57 | -0.3 | 9632.42 |
| 3 | epiforecasts-EpiExpert | 19 | 0.39 | 0.32 | 0.79 | 5589.31 | 2202.37 | 1466.14 | 1920.81 | -0.02 | 8785.37 |
| 4 | LANL-GrowthRate | 18 | 0.4 | 0.28 | 0.67 | 5562.97 | 1236.87 | 3832.15 | 493.95 | -0.36 | 7282.78 |
| 5 | EuroCOVIDhub-ensemble | 19 | 0.41 | 0.42 | 0.95 | 5499.53 | 2192.66 | 2650.68 | 656.19 | -0.29 | 8653.32 |

# Paper III - Discussion

- Still hard to generalise due to low statistical power (e.g. comparison between 'experts' and 'non-experts')
- The period chosen seems to be a period of constant exponential growth

  → it may be sensible to extend the study period
- Hybrid forecasting results very much depend on the specific implementation

# Paper III - Current progress

- Ethics approval received
- Data collection is in progress

# Paper IV

Ensemble sizes and optimal ensembles in epidemiological forecasting

- Background and motivation
- Aim and objectives
- Methods
- Discussion
- Current progress

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

# Paper IV - Background and motivation

- For both human forecast studies, we have submitted an ensemble of predictions to the Forecast Hubs, BUT: Unclear, which ensemble type is best suited
- Previous studies (Brooks et al. 2021): Median tends to be better than mean ensemble

  → is that true for all ensemble sizes? Is it true for human forecasts?
- Some of our previous forecasts improved the German and Polish Forecast Hub ensemble, some did not

  → when do models make a positive contribution? Is it worth to submit the 50th model?

# Paper IV - Aim and objectives

**Aim**: to obtain an understanding of how the choice of an optimal ensemble depends on the characteristics and the number of available models

**Objectives**:

- to analyse performance of different ensemble types for differing ensemble sizes
- to identify characteristics of ensembles that perform well
- to analyse when models make a positive or negative contribution to an existing ensemble

# Paper IV - Methods

- Use data from Forecast Hubs and our crowd forecasts
- For different ensemble sizes *n*, sample *n* forecasts and combine them to an ensemble. Analyse performance of different ensemble types depending on *n*
- Analyse characteristics of ensembles that do well (e.g. how similar are its component models? → Cramér-distance (Bellemare et al. 2017)
- Use 'leave-one-out'-ensembles to determine the value of each model's individual contribution to an ensemble

# Paper IV - Discussion

- The value that a model provides to an existing ensemble may depend on the type of ensemble (e.g. mean vs. median)
- The Cramér-distance only identifies how similar model *outputs* are, not how different model assumptions are

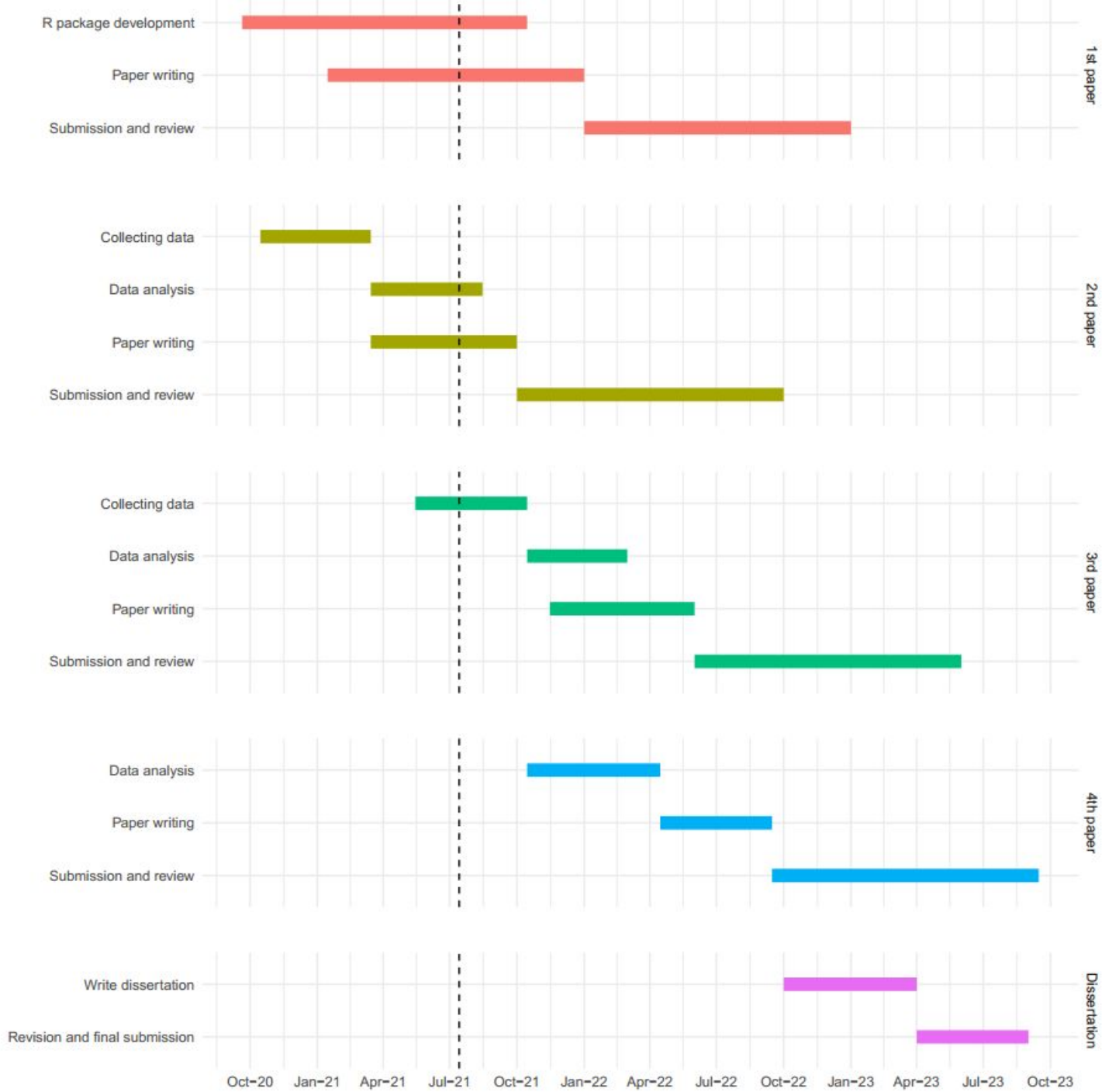
  → if two very different models come to similar conclusions, this should increase our confidence in this conclusion in a way not captured here

# Paper IV - Current progress

- Data is readily available
- No analysis has happened yet

# Timetable

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE




LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

# Thank you!

## Supervisors

Sebastian Funk (LSHTM)

Anne Cori (Imperial)

Edwin van Leuuwen (PHE)

## Advisory Committee

Sam Abbott (LSHTM)

Johannes Bracher (KIT)

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

# Thank you for your attention!

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

# References

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

# References


- Bracher, Johannes, Evan L. Ray, Tilmann Gneiting, and Nicholas G. Reich. 2021. "Evaluating Epidemic Forecasts in an Interval Format." PLoS Computational Biology 17 (2): e1008618. https://doi.org/10.1371/journal.pcbi.1008618
- Gneiting, Tilmann, Adrian E. Raftery, Anton H. Westveld, and Tom Goldman. 2005. "Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation." Monthly Weather Review 133 (5): 1098–118. https://doi.org/10.1175/MWR2904.1.
- Gneiting, Tilmann, Fadoua Balabdaoui, and Adrian E. Raftery. 2007. "Probabilistic Forecasts, Calibration and Sharpness." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69 (2): 243–68. https://doi.org/10.1111/j.1467-9868.2007.00587.x.
- Gneiting, Tilmann, and Adrian E Raftery. 2007. "Strictly Proper Scoring Rules, Prediction, and Estimation." Journal of the American Statistical Association 102 (477): 359–78. https://doi.org/10.1198/016214506000001437.
- Logan C. Brooks, Evan L. Ray, Jacob Bien, Johannes Bracher, Aaron Rumack, Ryan J. Tibshirani, and Nicholas G. Reich. n.d. "Comparing Ensemble Approaches for Short-Term Probabilistic COVID-19 Forecasts in the U.S. - International Institute of Forecasters." Accessed July 12, 2021. https://forecasters.org/blog/2020/10/28/comparing-ensemble-approaches-for-short-term-probabilistic-covid-19-forecasts-in-the-u-s/.
- Abbott, Sam, Joel Hellewell, Joe Hickson, James Munday, Katelyn Gostic, Peter Ellis, Katharine Sherratt, et al. 2020. "EpiNow2: Estimate Real-Time Case Counts and Time-Varying Epidemiological Parameters." https://doi.org/10.5281/zenodo.3957489.
-

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

# Additional slides

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

# Other project ideas

- Post-processing human forecasts to mitigate overconfidence

  → can we multiply uncertainty with a given factor?

- Evaluation of Rt forecasts against their own hindsight estimate to evaluate internal consistency of forecasts

- Looking into how different baseline models change crowd forecasts

# Conferences

- International Symposium on Forecasting 2022: Oxford, UK | July 10-13
- UserR conference 2022
- CBMS 2022 - Regional Conference in the Mathematical Sciences: Bayesian Forecasting and Dynamic Models (https://cbms.soe.ucsc.edu/home/cbms-2022)

# Training courses

Potential Moodle courses

- Epidemiology of Infectious Diseases (https://www.lshtm.ac.uk/media/41606)
- Machine Learning (https://www.lshtm.ac.uk/media/41656)
- Generalized Linear Models (https://www.lshtm.ac.uk/media/41631)
- Adv. Statistical Methods in Epidemiology (https://www.lshtm.ac.uk/media/41671)