PhD Upgrading report

Nikos Bosse

Department of Infectious Disease Epidemiology
Faculty of Epidemiology and Population Health
Centre for Mathematical Modelling of Infectious Diseases

London School of Hygiene & Tropical Medicine

July 2021

Supervisors

Sebastian Funk

Professor of Infectious Disease Dynamics and Wellcome Trust Senior Research Fellow

London School of Hygiene & Tropical Medicine

Anne Cori

Lecturer in Outbreak analysis and modelling, Faculty of Medicine, School of Public Health

Imperial College London

Edwin van Leeuwen

Senior Mathematical Modeller

Public Health England

Advisory Committee

Sam Abbott

Research Fellow in Real-time Modelling

London School of Hygiene & Tropical Medicine

Johannes Bracher

Postdoctoral Researcher

Chair of Statistics and Econometrics at Karlsruhe Insitute for Technology

Abbreviations

Abbreviation	Meaning
CFR	Case fatality ratio
CRPS	Continuous ranked probability score
DSS	Dawid-Sebastiani score
logS	Log score
MADN	Median absolute deviation about the median
PIT	Probability integral transform
WIS	Weighted interval score

Contents

1 Abstract

Background

Infectious disease modelling and forecasting can play a critical role in informing public health policy, as was highlighted by the COVID-19 pandemic. To improve infectious disease forecasting, learning from past predictions is indispensable. This requires appropriate tools to evaluate predictions, which can be used to obtain a deeper understanding of different forecasting approaches. Infectious disease forecasts are usually not only informed by model-based assumptions, but also implicitly by the opinion of the researchers implementing a model. This interplay between human judgement and model-based inference has not been studied in detail. Studying relative strengths of human forecasts and model-based approaches, and how to combine them, may therefore yield important insights. To further enhance predictive accuracy, individual predictions, be it from humans or models, are usually combined into ensembles. However, especially for small ensembles, it is not clear which aggregation method should be used and in which circumstances a model should be added to an existing ensemble or left out.

Aim

The aim of this PhD is to improve infectious disease forecasting and its usefulness to public health officials in the UK and other countries.

Objectives

1. Establish appropriate tools to evaluate predictions and ummarise best practices in forecast evaluation.
2. Collect predictions of COVID-19 from humans in Germany, Poland. Compare these human predictions against model-based forecasts to discern relative strengths and weaknesses of human forecasters vs. model-based approaches
3. Collect human forecasts of reported cases and deaths from COVID-19 in the UK as well as human predictions of the effective reproduction number R_t to explore ways in which human insight and epidemiological modelling can be combined
4. Examine how different numbers of forecasts can best be combined to model ensembles and identify circumstances in which individual models contribute most to those ensembles.

2 Introduction and aims

2.1 Role of infectious disease forecasting

Accurate knowledge of the future is immensely valuable. Good forecasts therefore are of great interest to decision makers in a multitude of fields like finance, weather predictions or infectious disease modeling (Funk et al. 2020). Model based forecasts of infectious diseases have a rich history and have been growing in popularity over the last decade (McGowan et al. 2019; Johansson et al. 2019; Viboud et al. 2018; Funk et al. 2019). Improving our understanding of what a good forecast is and how to make better predictions is an aim that is worth pursuing and can potentially have a large and lasting impact on public health decision making. The COVID-19 pandemic has once more underlined the importance of accurate infectious disease forecasting. It also highlighted the role of two topics closely related to forecasting: forecast evaluation and forecast aggregation. Modelling by influential research groups (Ferguson et al. 2020; IHME COVID-19 health service utilization forecasting team and Murray 2020) was impactful on policy decisions early in the pandemic, despite previous work having shown that relying on a single model can lead to less accurate forecasts than decisions based on multiple approaches (Yamana, Kandula, and Shaman 2016; Gneiting and Raftery 2005). Researchers and their forecasts were often criticised for a lack of accountability as predictions were rarely evaluated and compared against actual observations. Since then several collaborations have sought to improve COVID-19 forecasting by eliciting submissions from a large number of research teams and collecting them in forecast hubs in the United Kingdom (Funk et al. 2020), in the United States of America (Cramer et al. 2020; Cramer et al. 2021), in Germany and Poland (Bracher, Wolffram, et al. 2021), and in Europe (ECDC 2021).

2.2 Evaluating epidemiological forecasts

Model evaluation is an integral of the forecasting process that can provide valuable insights. It can help to choose between different models, but also provide a better understanding of how a model works and how it can be improved. One central aspect of forecast evaluation is the forecasting paradigm (Gneiting et al. 2005; Gneiting, Balabdaoui, and Raftery 2007) which states that a forecaster should aim to maximise the *sharpness* of their forecast subject to *calibration*. Sharpness is a feature of the forecast only and refers to how narrow or wide a prediction is. Calibration refers to the statistical consistency between the predictive distribution and the observations. Maximising sharpness subject to calibration therefore means that the goal is to have a forecast that is as precise as possible while still correct. Forecast performance is usually summarised using proper scoring rules, i.e. metrics which cannot be cheated and which make sure that a forecaster always states their best belief ((Bracher, Ray, et al. 2021; Gneiting, Balabdaoui, and Raftery 2007; Gneiting and Raftery 2007)). Other additional approaches have been employed in epidemiological settings to analyse specific aspects of the forecasts more closely (Funk et al. 2019; Cramer et al. 2021). While the literature on different evaluation metrics is extensive, actually conducting a forecast evaluation is difficult in practice due

to a lack of comprehensive guidelines and available tools. The first aim of this PhD is therefore to summarise and expand on best practices existing in the literature, as well as provide comprehensive and easy to use tools for forecast evaluation. This will form the basis on which the later chapters can build.

2.3 The role of human insight in infectious disease forecasting

Over the past months, thousands of model-based forecasts have been submitted to various COVID-19 forecasting hubs (Cramer et al. 2020; Cramer et al. 2021; Bracher, Wolffram, et al. 2021; ECDC 2021). These models in turn have been influenced by the researchers who adapted and tuned the models. The resulting predictions therefore are usually an implicit combination of the researcher's subjective opinion and model assumptions. Thinking about forecasts as existing on a spectrum between human opinion and model-based assumptions is a perspective which has not garnered much attention in the past. It is helpful, because it allows us to better understand aspects of forecasting where humans are good and those where predominantly model-guided predictions excel. A variety of human expert elicitation as well as crowd forecasting projects exist (McAndrew et al. 2021; Metaculus 2020; Tetlock et al. 2014; Atanasov et al. 2016). However, these crowd forecasts were not designed to be compared against model derived forecasts and usually follow a different (often binary) format or focus on more nuanced questions. In addition, no tools were available that would allow for a direct comparison of human prediction and model-based forecasts. The second aim of this thesis is therefore to elicit human predictions that can be directly compared with forecasts purely based on epidemiological modelling in order to examine relative strengths and weaknesses that may help improve infectious disease forecasting. To that end I have created an `R shiny` app to collect human predictions. These forecasts have been submitted to the German and Polish Forecast Hub (Bracher, Wolffram, et al. 2021) alongside other model-based predictions against which they can be compared.

Results obtained so far from the forecasts submitted to the German and Polish Forecast Hub suggest that an ensemble of human forecasters is able to predict future reported cases very well, but performs relatively worse at forecasting deaths. One possible hypothesis is that humans are relatively good at anticipating future changes in conditions (e.g. differing behaviour, environmental conditions or future interventions) that are hard to encode in a predictive computational model. On the other hand, they potentially struggle with quantifying the delays between observed cases and reported deaths which model-based forecasts may do better. These results, however, are subject to limitations as the number of participants was quite small and I only analysed human predictions on an aggregate level (an ensemble of human forecasts was analysed rather than individual predictions). In order to confirm (or reject) the patterns observed, I started the UK COVID-19 Crowd Forecasting Challenge that collects human predictions of reported cases and deaths in the UK from 24/05/2021 to 16/08/2021. Using this data will allow to gain additional insights made possible through a larger sample size. Especially, it is of interest whether a larger number of participants improves results, whether expert knowledge makes a difference and how individual forecasters, as opposed to an aggregate ensemble, perform. In

addition, a second forecasting method is tested, where (instead of a direct forecast) participants can make a forecast of the effective reproduction number R_t that gets then mapped to observed cases and deaths. The third aim is therefore to obtain a better understanding of how individual humans predict COVID-19 and whether their predictions can be enhanced by using a hybrid approach that makes use of epidemiological insights.

2.4 Improving epidemiological forecasts by means of ensembling

Single predictions can be combined into ensembles. One approach, for example, is to take the average of all predictions and use that as the combined forecasts, thereby forming a unweighted mean ensemble. Other approaches, like taking the median instead of the mean or using a weighted instead of an unweighted average, are possible. In the past, ensemble-based approaches often have led to superior performance when compared to single model forecasts (Yamana, Kandula, and Shaman 2016; Gneiting and Raftery 2005). Understanding forecast ensembles is therefore crucial in order to improve infectious disease forecasting. Past research has shown that it is difficult to improve on equal-weighted ensembles (Claeskens et al. 2016). Current efforts associated with the US Forecast Hub investigate different forms of trained ensembles and compare them to untrained ensembles, showing promising results. One important aspect that has been neglected so far is the dependence of the optimal ensemble on the number of available ensemble members. Especially for smaller ensembles, common in many public health settings (Funk et al. 2020), good understanding of the relation between ensemble performance and size is important. In addition it is interesting to analyse what different types of models can add to an ensemble to understand in which situations adding a model to an ensemble may be beneficial or harmful. The fourth aim of the PhD thesis is therefore to gain a deeper understanding of the relation between ensemble size and performance for different ensemble types as well as the contributions that individual models make to these ensembles.

2.5 Aims and objectives

The aim of this PhD is to improve infectious disease forecasting and its usefulness to public health officials in the UK and other countries. In particular, it aims to address three key questions that pertain to different aspects crucial to infectious disease forecasting. The first one is: How can forecasts best be evaluated, in order to learn the most from past forecasts and improve accuracy of future predictions? The second one is: What role should human insight, as opposed to purely model-based inference, play in infectious disease forecasting and how can we best combine the two? The third one is: How can we best combine different predictions into a single forecast and how does the choice of an optimal aggregation method depend on the number and characteristics of available ensemble models?

Its first objective is to summarise best practices in forecast evaluation and to establish appropriate tools to evaluate predictions in `R`, which will be used to evaluate short-term forecasts of COVID-19 in the UK. In order to learn more about how humans. Secondly, human predictions of COVID-19 in Germany, Poland will be collected using an self-developed `R shiny` app. These predictions will

be compared against model-based forecasts to discern relative strengths and weaknesses of human forecasters and model-based approaches. In order to explore ways in which human insight and epidemiological modelling can be combined, human forecasts of reported cases and deaths from COVID-19 in the UK as well as human predictions of the effective reproduction number R_t will be collected. From the R_t forecasts, hybrid predictions will be obtained by mapping R_t to reported cases and deaths using the renewal equation, which can then be compared against direct predictions. Lastly, in order to examine how different numbers of forecasts can best be combined to model ensembles, data previously collected will be combined using different ensembling techniques and properties of these ensembles will be studied.

3 Evaluating epidemiological forecasts - tools and best practices (Paper 1)

3.1 Aim and objective

The first aim of this PhD is to establish appropriate tools to evaluate predictions in `R` and summarise best practices in forecast evaluation. The `scoringutils` package, which I have developed, makes numerous scoring metrics and proper scoring rules available in a coherent framework. The first chapter of my PhD will summarise the metrics available in the `scoringutils` package, discuss best practices and apply the tools to an evaluation of short-term forecasts of COVID-19 (Funk et al. 2020). The `scoringutils` package as well as the discussion of best practices in forecast evaluation will form the foundation on which later chapters can build.

3.2 Introduction

Evaluating past forecasts is indispensable to assess and improve the accuracy of predictions for the future and is therefore of great interest in public health policy making. For decades, researchers have developed and refined an arsenal of techniques not only to forecast, but also to evaluate these forecasts (see e.g. Bracher, Ray, et al. (2021), Funk et al. (2019), Gneiting, Balabdaoui, and Raftery (2007), and Gneiting and Raftery (2007)). Yet even with this rich body of research available, implementing a forecast evaluation in is not trivial.

The first reason for this is a lack of adequate and easy to use tooling. Some `R` packages exist that bundle different scoring metrics together, but none offer the user a standalone solution to forecast evaluation. The `scoringRules` package (Jordan, Krüger, and Lerch 2019) offers a very extensive collection of proper scoring rules. However, its implementation is very technical and it lacks important features needed in the evaluation process. Other packages like `Metrics` (Hamner and Frasco 2018) and `MLmetrics` (Yan 2016) are geared towards machine learning problems and don't implement the set of metrics and scoring rules desired for forecast evaluation. Secondly, the multitude of available methods published across various papers can make it difficult to obtain a comprehensive overview of which metric to use and how to interpret the results.

In order to address this, I have developed the `scoringutils` package. The package and the accompanying paper provides users with the tools as well as the necessary knowledge to conduct a thorough forecast evaluation and interpret the results. The `scoringutils` package brings forth a standardised and tested toolkit. It offers convenient automated forecast evaluation in a data.table format, but also provides experienced users with a set of reliable lower-level scoring metrics they can build upon in other applications. In addition it implements a wide range of flexible plots that are able to cover most day-to-day use cases. The paper provides an overview of the fundamental ideas behind forecast evaluation, gives a detailed explanation of the evaluation metrics in and discusses what needs to be considered when applying them in practice. It then presents a case study based on the evaluation of

COVID-19 related short-term forecasts in the UK (Funk et al. 2020).

3.2.1 Forecast types and forecast formats

In its most general sense, a forecast is the forecaster's stated belief about the future (Gneiting and Raftery 2007) that can come in many different forms. Quantitative forecasts are either point forecasts or probabilistic in nature and can make statements about continuous, discrete or binary outcome variables. Point forecasts only give one single number for the most likely outcome, but do not quantify the forecaster's uncertainty. This limits their usefulness, as a very certain forecast may, for example, warrant a very different course of actions than does a very uncertain one. Probabilistic forecasts, in contrast, by definition provide a full predictive distribution. This makes them much more useful in any applied setting, as we learn about the forecaster's uncertainty and their belief about all aspects of the underlying data-generating distribution (including e.g. skewness or the width of its tails) (Held, Meyer, and Bracher 2017). Probabilistic forecasts are therefore the focus of the paper as well as the package. The predictive distribution of a probabilistic forecast can be represented in different ways with implications for the appropriate evaluation approach. The paper elaborates on the different forecast types and when to use which scoring metrics.

3.2.2 The forecasting paradigm

Any forecaster should aim to minimise the difference between the (cumulative) predictive distribution F and the unknown true data-generating distribution G (Gneiting, Balabdaoui, and Raftery 2007). For an ideal forecast, we therefore have

$$F = G,$$

where F and G are both cumulative distribution functions. As we don't know the true data-generating distribution, we cannot assess the difference between the two distributions directly. (Gneiting, Balabdaoui, and Raftery 2007) instead suggest to focus on two central aspects of the predictive distribution, calibration and sharpness (illustrated in Figure 1. Calibration refers to the statistical consistency (i.e. absence of systematic deviations) between the predictive distribution and the observations. Sharpness is a feature of the forecast only and describes how concentrated the predictive distribution is, i.e. how precise the forecasts are. The general forecasting paradigm states that we should maximise sharpness of the predictive distribution subject to calibration. A model that made very precise forecasts would at best be useless if the forecasts were wrong most of the time. On the other hand, a model may be well calibrated, but not sharp enough to be useful. Take a weather forecast that would assign 30 percent rain probability for every single day. It may be (marginally) calibrated when looking at the average rainfall over the course of a year, but it doesn't give much guidance on a day to day basis. (Gneiting, Balabdaoui, and Raftery 2007) discuss different forms of calibration in more detail.

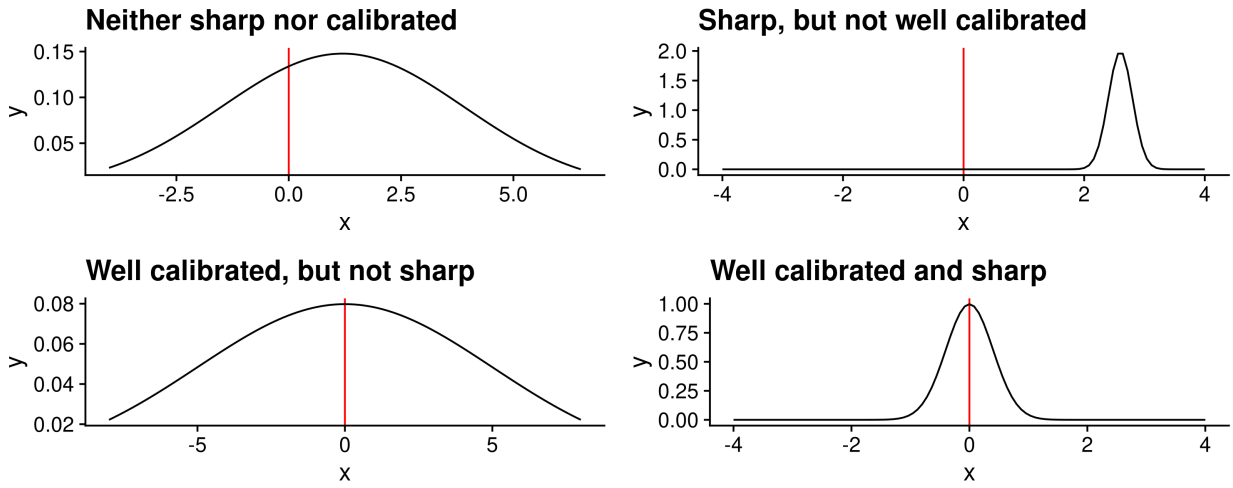


Figure 1: Schematic illustration of calibration and sharpness. True value are represented in red, the predictive distribution is shown in black.

3.3 Evaluation metrics and evaluation approaches

Some of the metrics in `scoringutils` focus only on sharpness or on calibration. Others, called proper scoring rules, combine both aspects into a single number. The former can be helpful to learn about specific model aspects and improve them, the latter are especially useful to assess and rank predictive performance of a forecaster. Evaluating calibration and sharpness independently is helpful for model diagnostics. To that end `scoringutils` makes numerous metrics available that aim to capture different aspects of sharpness and calibration.

Proper scoring rules (Gneiting and Raftery 2007) jointly assess sharpness and calibration and assign a single numeric value to a forecast. A scoring rule is proper if a perfect forecaster (the predictive distribution equals the data-generating distribution) receives the lowest score on average. This makes sure that a forecaster evaluated by a proper scoring rule is always incentivised to state their best estimate. The most important proper scoring rules are the continuous ranked probability score (CRPS), the log score (logS), the weighted interval score (WIS) and the Dawid-Sebastiani score (DSS). Often, the type of the forecasts restricts the use of the scoring rule. Where this is not true, different scoring rules involve different trade-offs which the paper discusses in detail.

3.3.1 Evaluating short-term forecasts in the UK

The metrics implemented in `scoringutils` will be used to evaluate short-term targets of healthcare targets such as the number of hospitalisations and deaths in the UK using the data from Funk et al. (2020).

Metric	Target types	Forecast formats	Properties
(Continuous) ranked probability score (CRPS)	continuous, discrete	closed-form, samples (approximation)	proper scoring rule, global, stable handling of outliers
Log score (logS)	continuous, (discrete not in scoringutils)	closed-form, samples (approximation)	proper scoring rule, local, unstable for outliers
(Weighted) interval score (WIS)	continuous, discrete	quantile or interval predictions	proper scoring rule, global, stable handling of outliers, converges to crps
Dawid-Sebastiani score (DSS)	continuous, discrete	closed-form, samples (approximation)	proper scoring rule, somewhat global, somewhat stable handling of outliers
Brier score (BS)	binary	binary probabilities	proper scoring rule
Interval coverage	continuous, discrete	interval forecasts (needs matching quantiles)	measure for calibration
Quantile coverage	continuous, discrete	quantile or interval forecasts	measure for calibration
Probability integral transform (PIT)	continuous, discrete, quantile	closed-form, samples, quantile or interval forecasts	assesses calibration
Sharpness	continuous, discrete	closed-form, samples, quantile or interval forecasts	measures sharpness, slightly different depending on forecast format
Bias	continuous, discrete, quantile	closed-form, samples, quantile or interval forecasts	captures tendency to over-or underpredict (aspect of calibration)
Mean score ratio	depends on score	depends on score	compares performance of two models
Relative skill	depends on scored	depends on score	Ranks models based on pairwise comparisons

Figure 2: Overview of the scoring metrics implemented in scoringutils.

3.4 Current progress

The `scoringutils` package itself is operational and on CRAN. All major functions are unit tested. Before publication, a few edits still need to be made, especially with regards to plotting functionality. A first draft of the paper is written that includes a detailed description of all scores and explains when to use them and how to interpret the results. Figures and illustrations for the paper need to be reworked.

4 The role of human insight in epidemiological modelling - comparing crowd forecasts and model based predictions of COVID-19 (Paper 2)

4.1 Aim and objective

The second aim of my PhD is to obtain a better understanding of what role should human insight, as opposed to purely model-based inference, play in infectious disease forecasting. The second chapter of my PhD presents a study in which human forecasts of COVID-19 in Germany and Poland have been collected through a self-developed `R shiny` app and submitted to the German and Polish Forecast Hub alongside two untuned model-based forecasts. By comparing the ensemble of human predictions against model-based forecasts, relative strengths and weaknesses are analysed.

4.2 Introduction

Over the last months, several collaborations have sought to improve COVID-19 forecasting by eliciting submissions from a large number of research teams and collecting them in forecast hubs in the United Kingdom (Funk et al. 2020), in the United States of America (Cramer et al. 2020; Cramer et al. 2021), in Germany and Poland (Bracher, Wolffram, et al. 2021), and in Europe (ECDC 2021). Whilst all of these efforts have successfully delivered more accurate forecasts to policy makers compared to individual forecasting efforts they have struggled to unpick what leads to good COVID-19 forecasts (Cramer et al. 2021; Bracher, Wolffram, et al. 2021; Funk et al. 2020). This has been partly driven by the complexity of the models used to produce the constituent forecasts but also because of the level of expert intervention in most forecasting methods over time due to changes in the pandemic, and the available data. These issues can potentially be decoupled by separating infectious disease forecasting into model derived forecasts, that are unadjusted during the forecast period, and human elicitation forecasts (from now on referred to as crowd forecasts).

This work aims to explore the role of human insight by explicitly comparing an ensemble of human insight with forecasts derived from two epidemiological motivated models that we did not alter throughout the forecast period and an ensemble of models from other researchers which is likely to have been modified based on opinion. All forecasts were produced and submitted in real-time to the German and Polish Forecast Hub over 21 weeks from the 12th October 2020 to March 1st 2021 and combined, along with other forecasts, into an ensemble used by policy makers as well as being independently evaluated by the research group running the German and Polish Forecast Hub.

4.3 Methods

We submitted three different forecasts of reported and cases in Germany and Poland to the German and Polish Forecast Hub (Bracher, Wolffram, et al. 2021) between October 12th 2020 and March 1st 2021. The first of these was an ensemble of crowdsourced opinion. We compared this approach with two open-source real-time methods which we did not alter throughout the study period. The

first of these, the "renewal model," estimated the target observation by reconstructing infections using an autoregressive approach with the weighting based on the generation time between infections and then using a discrete convolution to estimate reported observations. The second approach, the "convolution model," assumed that a target observation, such as deaths, was a convolution of cases multiplied by a scaling factor.

4.3.1 Crowd forecast

Participants were asked to make forecasts of COVID-19 cases and deaths over a four week ahead horizon using a web application (https://cmmid-lshtm.shinyapps.io/crowd-forecast/). The application was built using the `shiny` and `golem` R packages (Chang et al. 2021; Fay et al. 2021) and is available in the `crowdforecastr` R package (N. I. Bosse et al. 2020). To make a forecast in the application participants could select a predictive distribution, with the default being log-normal, and adjust the median and the width of the uncertainty by either interacting with a figure showing their forecast or providing numerical values. The baseline shown was a repetition of the last known observation with constant uncertainty around it computed as the standard deviation of the last four observed log changes in forecasts. Information on the chosen distribution as well as the parameters for median and width were used to obtain a set of 22 quantiles plus the median from that distribution. Forecasts from all forecasters were then aggregated using an unweighted quantile-wise mean.

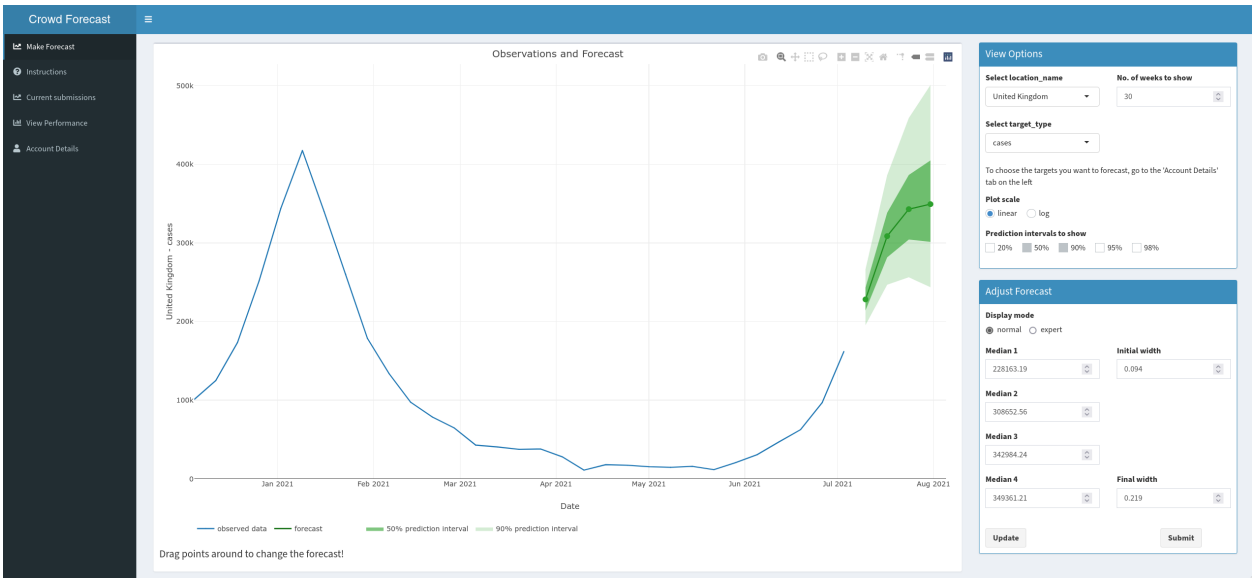


Figure 3: Screenshot of the crowdforecastr interface for the purpose of this PhD.

4.3.2 Forecast submission

Crowd predictions for Germany and Poland were collected every week up to a 4 week time horizon and submitted to the German and Polish Forecast Hub alongside the two model-based forecasts every Tuesday 3pm. The model based forecasts used data up to the previous Sunday. Human forecasters

were allowed to make forecasts until Tuesday 12am, but were asked to use only information up to Monday. All forecasts were submitted in a quantile-based format with 22 quantiles plus the median prediction for a one to four week ahead horizon.

4.3.3 Statistical analysis

Forecasts were analysed by visual inspection as well using the following scoring metrics: The weighted interval score (WIS) (Bracher, Ray, et al. 2021), absolute error, bias, and empirical coverage of the 50% and 90% prediction intervals. In addition to the WIS, we also calculated WIS relative to the ensemble of all other models submitted to the German and Polish Forecast Hub (rel.WIS). All scores were calculated using the `scoringutils` R package (N. Bosse 2020). For the main analysis we focused on two week ahead predictions, as predictions beyond this horizon are often unreliable due to rapidly changing condition (Bracher, Wolffram, et al. 2021). At all stages of the evaluation our forecasts were compared to the median ensemble of all other models submitted to the German and Polish Forecast Hub (hub-ensemble). In addition to this we assessed the impact of our forecasts on the realised performance of the forecasting hub by recalculating the hub-ensemble after including each of our forecasts in turn.

4.4 Results

4.4.1 Forecast submission

A total number of 31 participants submitted forecasts. The median number of forecasters per week was 6, the minimum 2 and the maximum 9. Participation rose steadily and peaked in February, before declining towards the end of the study period. The mean number of submissions from an individual forecaster was 4.7 but the median number was only one - most participants dropped out after their first submission. Only two participants submitted a forecast every single week both.

4.4.2 Performance overview

We found that crowd forecast had a lower mean WIS than the renewal model across all forecast targets, horizons and locations with a mean WIS for two week ahead predictions relative to the hub ensemble of 89% (crowd forecasts) and 140% (renewal model) for cases and 126% vs 179% for deaths (Figure 4). The convolution model forecast deaths better on average than the crowd forecast up to two weeks ahead (rel. WIS of 122% vs 126%), where deaths were largely informed by observed cases. It did less well on average at greater forecast horizons (rel. WIS of 180% four weeks ahead vs. 117%). The renewal model generally performed poorly at predicting deaths.

In comparison, using the median WIS, we found that the renewal model outperformed all other forecasts at the one week horizon across all targets and locations (Figure 4). However, as for the mean WIS this performance degraded rapidly as the horizon increased. Performance in comparison to other forecasts was relatively unchanged for the convolution model. The crowd forecast performed

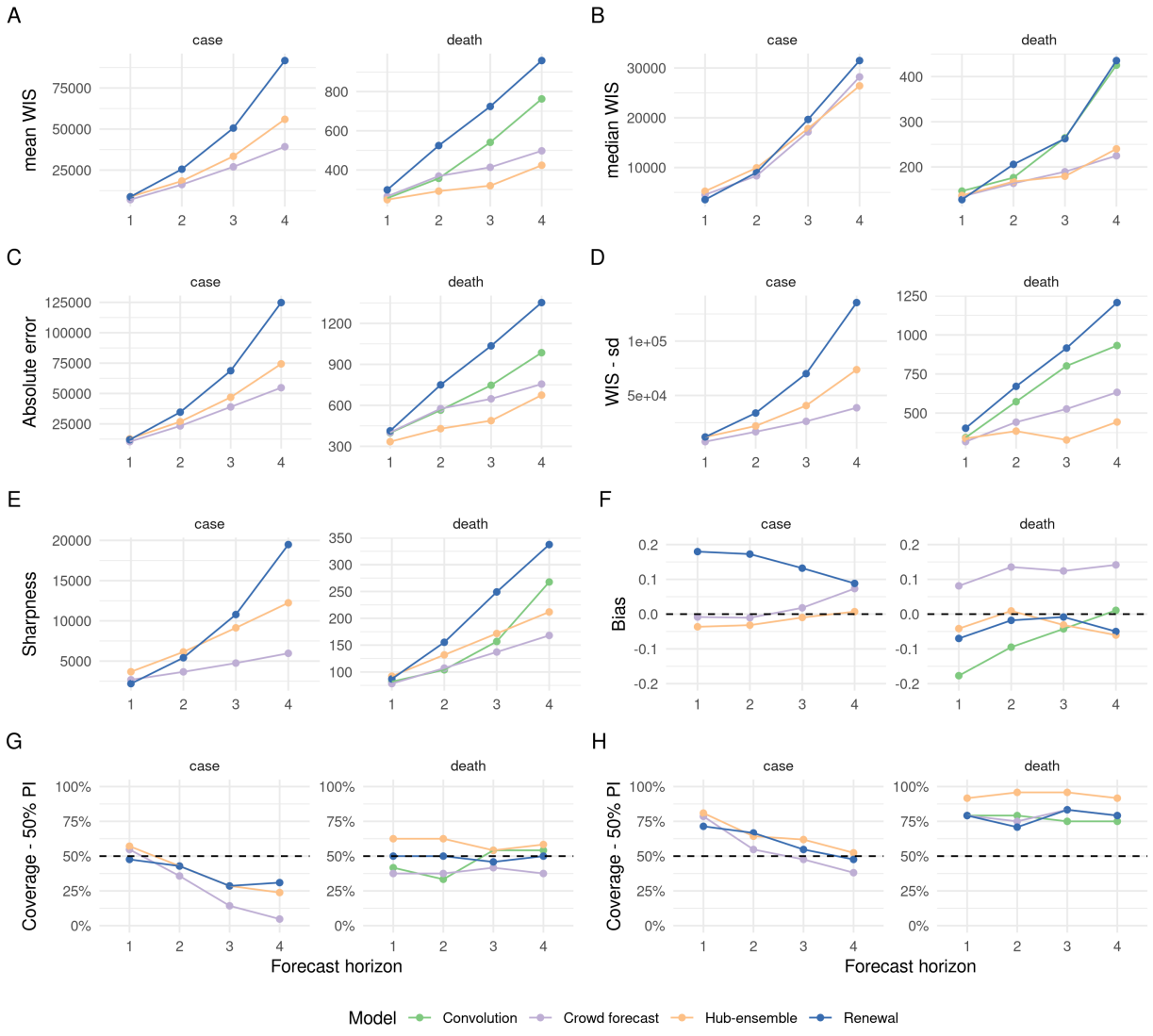


Figure 4: Visualisation of aggregate performance metrics across forecast horizons. A: mean weighted interval score (WIS) across horizons. B: median WIS. C: Absolute error of the median forecast. D: Standard deviation of the WIS. E: Sharpness (higher values mean greater dispersion of the forecast). F: Bias, i.e. general tendency to over- or underpredict. Values are between -1 (complete underprediction) and 1 (complete overprediction) and 0 ideally. G: Empirical coverage of the 50 percent prediction intervals. F: Empirical coverage of the 90 percent prediction intervals.

comparably or better than the hub-ensemble using the median WIS across all locations, targets and horizons.

Only the crowd forecast consisently out-performed the hub-ensemble when assessed by both median and mean WIS and forecasting cases. The hub ensemble performed better than all our forecasting approaches for forecasting deaths at longer time horizons when assessed using the mean WIS but performance was comparable using the crown ensemble when the median WIS was used. Our model based forecasts performed comparably to the hub-ensemble at short-time horizons but as noted performance rapidly degraded as the horizon increased.

4.5 Discussion and limitations

Submitting human crowd forecasts to a forecast hub expressly designed to evaluate and aggregate quantitative forecasts is a novelty and created a unique opportunity to directly and fairly compare human predictions against model-based forecasts as well as contribute to the forecasts available to public health policy makers. The findings shed light on potential structural patterns that distinguish human crowd forecasts, untuned model-based predictions and forecast models that are continuously improved by human intervention. They are, however, not directly generalisable.

First, our untuned models cannot represent all model-based forecasts. While we aimed to create two models that capture the simplest possible epidemiological baseline assumption about how an epidemic involves, these are still two particular models with particular strengths and weaknesses. Second, findings are confounded by the fact that we compared models and ensembles of models. Many of the features we observed, for example the ability or inability to avoid large outlier predictions, may be more a feature of ensembles, or the type of ensembles used here, than sign of any human intervention. Third, we were not able to directly observe the role of human insight in the models that were submitted to the German and Polish Forecast Hub. Fourth, while the methodology did not change for the renewal model and the convolution model, this continuity is not given for the crowd forecasts and the hub ensemble, as forecasters (or models, respectively), dropped in and out. Fifth, given the low number of participants, it is difficult to generalise conclusions about crowd predictions to other settings. In particular, our crowd forecasting application was relatively technical which may have precluded less technical, but interested parties, from submitting forecasts. It is both conceivable that a greater number of participants would have improved forecasts, but also that excluding a larger audience may have increased average quality of predictions.

4.6 Current progress

The study has received approval from the ethics committee of the London School of Hygiene & Tropical Medicine. The `crowdforecastr` app is in a functioning state and data collection for the study is completed. A first draft of the paper has been written and is currently awaiting feedback from Co-authors and supervisors.

5 The role of human insight in epidemiological forecasting - towards a deeper understanding (Paper 3)

5.1 Aim and objective

The third aim of this PhD is to learn more about how humans make predictions of COVID-19 and how human insight can best be combined with model-based inference. The third chapter will present a study which collects human predictions of reported cases and deaths from COVID-19 as well as forecasts of R_t in the UK. Going beyond the analysis of an ensemble of human forecasts presented previously, this chapter will examine individual predictions and analyse differences between individual forecasters. In addition, it will compare direct forecasts against R_t predictions in order to explore the potential for combining human insight with model-based approaches.

5.2 Introduction

Ongoing work with crowd forecasts in Germany and Poland suggests that human insight can play an important role in infectious disease forecasting, especially when predicting targets that strongly depend on factors which are hard to model such as future inventions or changes in behaviour. Model-based predictions, on the other hand, seem to be valuable when modelling targets such as the number of reported deaths that can be modelled using leading indicators such as cases or hospitalisations and knowledge of the delays between these. Previous work has looked at an ensemble of human forecasters which was compared with two untuned epidemiological baseline models as well as an ensemble of model-based, but expert informed, predictions. Looking at an ensemble of human predictions, however, masks the variability between forecasters. What made the ensemble of human forecasts successful in previous work therefore may to a large extent be a feature of ensembles, rather than a feature of human forecasts. Analysing the performance of individual forecasters is therefore important to draw more informative conclusions. A larger sample size and a different setting will help to confirm (or update on) the observations made in Germany and Poland. Based on past results, a second interesting question emerges: is there a way to combine the relative strengths of human forecasters and model-based approaches? To answer this question I developed a version of the `crowdforecastr` app that allows to make a forecast of the time-varying reproduction number R_t, rather than a direct forecast of reported cases and deaths. Using the renewal equation and a convolution, the estimate of R_t is then mapped to future cases and deaths.

5.3 Methods

Data on test positive cases and deaths linked to COVID-19 is provided by the organisers of the European Forecast Hub (ECDC 2021). Forecasts from different research institutions as well as an ensemble of all models submitted to the European Forecast Hub can be downloaded from the European Forecast Hub Github repository.

Crowd forecasts are collected over a period of 12 weeks from May 24th until August 16th 2021 as part

of a "COVID-19 UK Crowd Forecasting Challenge." Participants were recruited mainly by advertising the Forecasting Challenge on Twitter and with the help of a dedicated website, crowdforecastr.org. Participants were asked for a weekly submission of their forecast until 8pm UK time on Mondays. A submission could be made using two different R shiny apps with different underlying mechanics. Participants were explicitly allowed and encouraged to use both apps and submit two forecasts in order to increase their chances of winning.

The first app (https://cmmid-lshtm.shinyapps.io/crowd-forecast/) asked participants for a direct prediction of COVID-19 cases and deaths over a four week ahead horizon. To make a forecast in the application participants could select a predictive distribution, with the default being log-normal, and adjust the median and the width of the uncertainty by either interacting with a figure showing their forecast or providing numerical values. The baseline shown was a repetition of the last known observation with constant uncertainty around it computed as the standard deviation of the last four observed log changes in forecasts. For the direct forecast we required that participants submitted predictions with uncertainty that increased over time.

The second app (https://cmmid-lshtm.shinyapps.io/crowd-rt-forecast/) was an adaption of the original version that asked participants of a prediction for the time-varying reproduction number R_t. Forecasts were also elicited up to 4 weeks into the future, but as current R_t estimates are inherently uncertain ('nowcast'), participants were also asked to provide values for the past two weeks. Using a renewal equation as implemented in the R package `EpiNow2` (Abbott et al. 2020), the R_t trajectory was then mapped to future reported cases. Death forecasts were then obtained using a simple convolution model that predicted deaths as a convolution of observed and predicted cases with a delay distribution and a constant case fatality ratio (CFR) estimated from past observations. In the app, participants can simulate the case prediction that would result from a given R_t forecast, but cannot see the death forecast.

Both applications were built using the shiny and golem R packages (Chang et al. 2021; Fay et al. 2021) and are available in the crowdforecastr R package (N. I. Bosse et al. 2020). Our interface also allowed participants to view the observed data, and their forecasts, using a log scale and presented additional contextual COVID-19 data sourced from ("COVID-19 Data Explorer" n.d.). These data included notifications of both test positive COVID-19 cases and COVID-19 linked deaths, case fatality rates and the number of COVID-19 tests.

5.4 Results

5.4.1 Preliminary results

Results from this study are not yet in. Preliminary analysis shows that aggregate predictions from human forecasters so far have been very similar to the overall ensemble of all forecasts submitted to the Forecast Hub. I have succesfully recruited a larger number of participants than in the study we conducted in Germany and Poland, but recruiting participants is still difficult. We can see a

large heterogeneity in individual forecasts, with forecasters often predicting very different future trajectories. In particular, R_t forecasts are often quite different from direct predictions. Overall, we see far fewer participants for R_t forecasts than for direct forecasts.

5.4.2 Possible future results

Findings from the study in Germany and Poland suggest that humans are better at predicting cases than deaths. This could be either confirmed or rejected depending on the future findings. However, this analysis is complicated by the fact that the case fatality rate (CFR) in the UK is likely evolving over time with the roll-out of COVID-19 vaccines. Model-based forecasts submitted to the European Forecast Hub may have accounted for this or not and it is not clear how humans implicitly account for it.

Comparison of direct prediction vs. Rt forecast

We already see that R_t forecasts are sometimes different from the direct forecasts. This difference is expected for death forecasts, as participants cannot see the death forecasts that are generated from their Rt prediction. Death forecasts generated from R_t will perform either better, worse or similar to the direct forecasts. If R_t forecasts are better, this would suggest that the proposed hybrid forecasting approach works well. If they are worse, this would suggest that either hybrid forecasting does not work in general or that the the current implementation does not work. This could be because users did not understand the interface, because the assumption of a constant CFR did not hold, or because of other issues with the implementation. If R_t forecasts and direct prediction perform similarly, it will be interesting to analyse whether this is because forecasts are also similar, or whether there are systematic differences that yield similar results on average.

For case predictions, differences between direct and R_t forecasts are not expected, as as users can simulate and see the case forecasts their R_t predictions implies. It is therfore important to find out in which way forecasts differ and what drives differences in performance. There are three possible types of disagreements between forecasts that could occur:

- a shift (e.g. one forecasts is consistently higher or lower)
- a difference in the levels of uncertainty around the forecast
- a different shape of the forecast (a different trajectory is predicted)

A shift between forecasts may come from differences in the interface where it is hard to reproduce predictions exactly. A difference in uncertainty may come from the fact that R_t forecasts suggest a certain structure in the uncertainty and also from the model uncertainty that the renewal equation model enforces. A difference in the shape / trajectory may indicate that participants think differently about their prediction if asked to make it in the form of Rt. It may also indicate that users try to hedge their predictions. Given that everyone may submit two predictions it may make sense to submit differing forecasts. Generally, user error and a lack of understanding how the R_t forecast

works cannot be ruled out. It is interesting to see which forecasts are better and for what reason (e.g. level of uncertainty, better absolute error, i.e. predictions closer to observations)

Consistency in forecast performance

In order to improve forecasting it is important to know whether good forecasters are consistently good and whether past performance can predict future performance. Rankings could either stay constant or change a lot over time. One additional possible way to check this would be to randomly remove forecast dates from the evaluation and check how robust the rating is. To analyse systematic patterns, it will be interesting to look at whether individual forecasters constantly overpredict / underpredict or whether they are constantly over- / underconfident?. Also, an interesting question is how average performance correlates with the number of submissions from an individual and whether individuals who submitted both forecasts are on average better.

5.5 Discussion and limitations

Compared to the study done in Germany and Poland, the higher number of participants potentially allows for slightly stronger conclusions about the potential of an ensemble of human forecasters. However, the sample size is still too small and varies too much across weeks to draw confident conclusions. While we ask participants whether they are an 'expert' and work in epidemiology, it is unclear how conclusive that self-identified information is. Results of the hybrid forecasts need to be taken with care as the relationship between cases and deaths likely has changed over time due to increasing vaccination of the British population. In addition, forecasters were not able to control all aspects of the hybrid forecasts, as e.g. the renewal equation enforced some model uncertainty that participants could not change. Apart from this our study in the UK only represents forecasts for one very specific epidemic curve and potentially more countries would need to be included to check robustness.

Participants may have tried to game the scoring rule: For the UK COVID-19 Crowd Forecasting Challenge, Participants can submit two different predictions, and each of them is independently eligible for a prize. This, in effect, means that every person has two shots at winning the competition. This decision was taken in order to encourage participants to submit a forecast using both the direct forecasting app and the R_t app (while not forcing them to use both). Having two forecasts from the same individuals allows to make a more direct comparison between the two forecasts, yielding potentially more interesting results. However, this may also have incentivised participants to try and game the competition. Every individual forecast is scored using a proper scoring rule which incentivises the forecaster to submit their best possible prediction. If overall performance were taken as the average score of two model submissions, this property would hold for every individual forecast and participants would have been incentivised to submit their best, i.e. the same, prediction twice. In this scenario, however, the two forecasts represent two independent possibilities to win a prize. Submitting the same forecast twice may therefore not be the optimal strategy (consider someone who

could submit 500 forecasts). As forecasts scores for both approaches are averaged across twelve weeks (and so one would have to be lucky quite often), the actual potential to game the system is probably minimal. Nevertheless, forecasters may have believed that submitting two different forecasts may be beneficial ("to be right at least once"), confounding observed results.

5.6 Current progress

The study has received approval from the ethics committee of the London School of Hygiene & Tropical Medicine. The `crowdforecastr` app is in a functioning state and data collection for the study is in progress, with a median number of participants of 20.

6 Ensemble sizes and optimal ensembles in epidemiological forecasting (Paper 4)

6.1 Aim and objective

For the two previous studies, I submitted a mean ensemble of human predictions to the German and Polish as well as the European Forecast Hub. However, given the low number of participants (especially in the first study) and very heterogeneous predictions, it was unclear what method should be used to aggregate individual forecasts. The fourth aim of this PhD therefore is to obtain a better understanding about how different forecasts can best be combined into a single forecast in similar situations. In particular, it aims to analyse how the choice of an optimal aggregation method depends on the number and characteristics of available forecasts. This is partly motivated by the fact that the studies in the last two chapters have submitted ensembles of human predictions to the German and Polish Forecast Hub as well as the European Forecast Hub, without having a good understanding of which aggregation method would be the optimal choice. The fourth chapter of my PhD will examine this question by using data previously collected as well as forecasts submitted to the European Forecast Hub and combining predictions to ensembles of different sizes.

6.2 Introduction

Ensembles usually perform better at forecasting than individual models (Yamana, Kandula, and Shaman 2016; Gneiting and Raftery 2005). Collecting forecasts in the form of Forecast Hubs therefore is an important step towards improving infectious disease forecasting. Past research has shown that equally weighted ensembles perform very well in comparison to trained ensembles (Claeskens et al. 2016). Creating ensembles that are able to outperform a simple average of all member forecasts is difficult, but possible as ongoing research efforts within the Forecast Hubs suggest (Logan C. Brooks et al. n.d., n.d.). But even if deciding against a weighted ensemble, a choice has to be made whether forecasts shall be combined using a mean or a median. Past research has put little focus on how the optimal choice for an aggregation method depends on the number of available forecasts. In addition, whether or not a model makes a positive contribution to an ensemble may depend on the type of ensemble. For a median ensemble, for example, the direction of a forecast is more dependent than the magnitude of that direction, implying that even extreme and badly calibrated forecasts can make a positive contribution. This, however, may not be the case for mean ensembles. Analysing and identifying situations in which a model is likely to make a positive contributions is therefore important. The contribution a model makes to an ensemble may also depend on how similar this model is to other models in the ensemble. One possible way to assess this dissimilarity is the Cramér-distance between the forecasts of two models. A model's contribution will also presumably depend on the size of the existing ensemble. Whether or not to spend researcher time to contribute a model to an existing ensemble or to prioritise something else is a question worth exploring in public health settings where resources are constrained.

6.3 Methods

6.3.1 Data sources

Several potential data sources are available for this study:

- forecasts from the European Forecast Hub
- forecasts from the US Forecast Hub
- forecasts from the German and Polish Forecast Hub
- crowd forecasts collected for the UK crowd forecasting challenge

All data are readily available through public github repositories.

6.3.2 Analysis

To analyse how different ensemble types perform depending on the ensemble size n, we iteratively sample n models from all available models, combine them to ensembles, and evaluate ensemble performance for different n. Performance will mainly be evaluated using the weighted interval score (WIS), a proper scoring rule, as well as the empirical coverage of the 50% and 90% prediction intervals. For every ensemble type and given ensemble size n, an average score will be computed that allows to compare different ensembling procedures against each other for varying ensemble sizes. In order to analyse the robustness of the results, variance in performance of different ensemble types will also be taken into account.

For any given ensemble, similarity between member models will be assessed using the Cramér-Distance between all possible model pairs. Analysing the relationship between model (dis-)similarity and ensemble performance will allow for a better understanding of the effects of ensemble composition. In addition, the contribution of individual models to an ensemble will be assessed by creating leave-one-out ensembles, where ensemble performance is computed once with and once without a given model. This will hopefully help to identify situations in which adding a model to an ensemble is beneficial or harmful.

In the beginning I intend to start with simple mean and median ensembles, but the analysis could easily be extended to arbitrary types of ensembles. To check robustness, all steps could be repeated across different data sets.

6.4 Expected Results

I hope to identify a relation between ensemble size and optimal ensemble type as well as conditions under which adding a model to an ensemble is beneficial. In addition I hope to be able to point out characteristics of ensembles that lead to better average performance or increased robustness in performance. Whether or not a model makes a positive contribution to an ensemble may well depend on the ensemble type. For a median ensemble, it is only important that a model shifts the ensemble in the correct direction, whereas for a median the magnitute of that shift matters. It therefore seems

that it may be easier for a model to contribute to a median ensemble, as only the direction of the forecast with respect to the average forecast needs to be correct, not the absolute level.

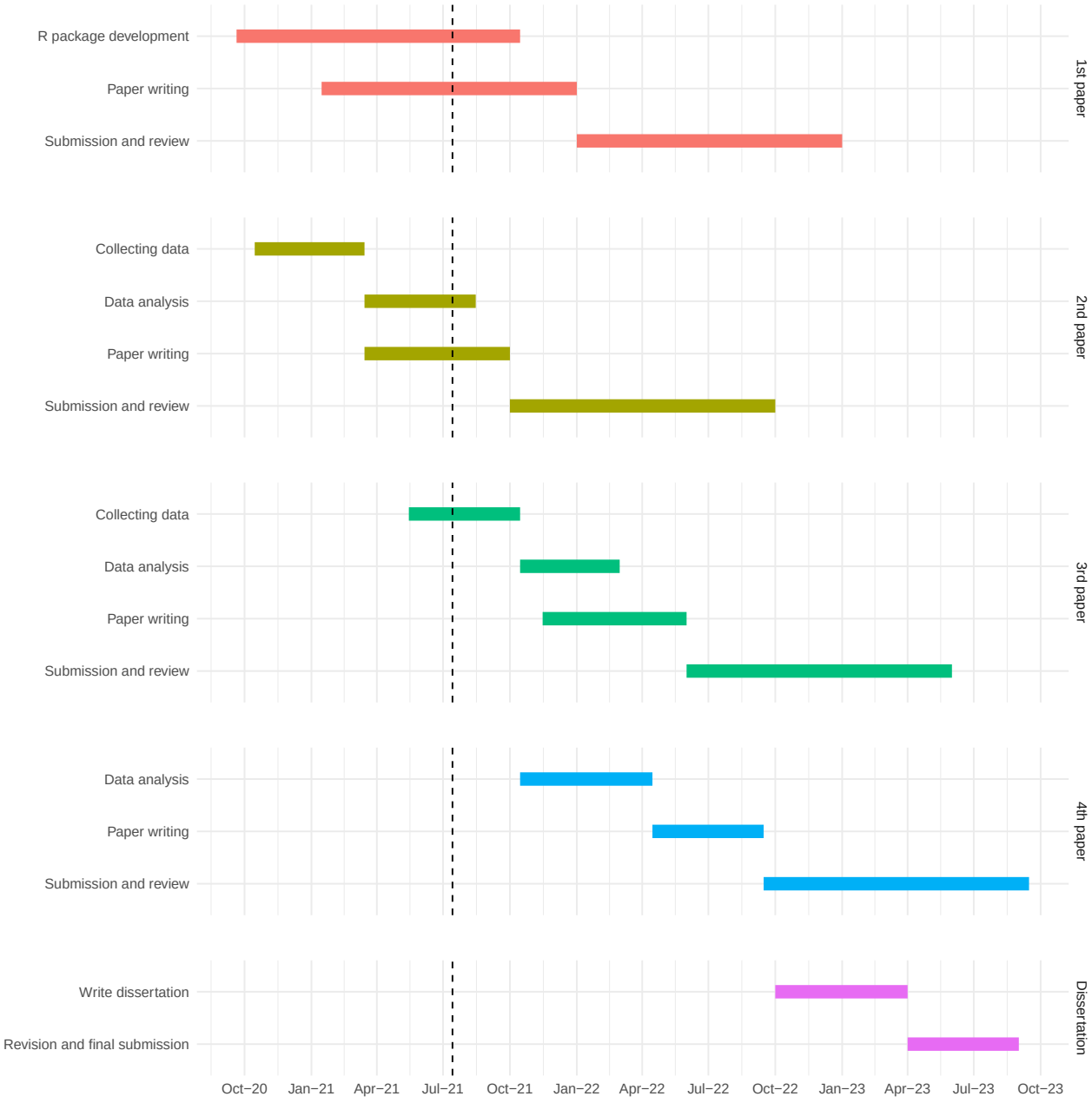
6.5 Limitations

The Cramér Distance can only measure how different two forecasts are, not how much model assumptions diverge. If two models with very different assumptions give similar answers than this should increase our confidence in the forecast in a way that is very hard to quantify. As in many applied settings model assumptions are very hard to assess, the Cramér-distance may however still provide useful information.

6.6 Current progress

Data is readily available, but work on this study has not yet started.

7 Proposed timetable



8 References

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.

Atanasov, Pavel, Phillip Rescober, Eric Stone, Samuel A. Swift, Emile Servan-Schreiber, Philip Tetlock, Lyle Ungar, and Barbara Mellers. 2016. "Distilling the Wisdom of Crowds: Prediction Markets Vs. Prediction Polls." *Management Science* 63 (3): 691–706. https://doi.org/10.1287/mnsc.2015.2374.

Bosse, Nikos. 2020. *Scoringutils: A Collection of Proper Scoring Rules and Metrics to Assess Predictions*. https://github.com/epiforecasts/scoringutils.

Bosse, Nikos I., Sam Abbott, EpiForecasts, and Sebastian Funk. 2020. *Crowdforecastr: Eliciting Crowd Forecasts in r Shiny*. https://doi.org/10.5281/zenodo.4618519.

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.

Bracher, Johannes, Daniel Wolffram, J. Deuschel, K. Görgen, J. L. Ketterer, A. Ullrich, S. Abbott, et al. 2021. "Short-Term Forecasting of COVID-19 in Germany and Poland During the Second Wave – a Preregistered Study." *medRxiv*, January, 2020.12.24.20248826. https://doi.org/10.1101/2020.12.24.20248826.

Chang, Winston, Joe Cheng, JJ Allaire, Carson Sievert, Barret Schloerke, Yihui Xie, Jeff Allen, Jonathan McPherson, Alan Dipert, and Barbara Borges. 2021. *Shiny: Web Application Framework for r*. https://CRAN.R-project.org/package=shiny.

Claeskens, Gerda, Jan R. Magnus, Andrey L. Vasnev, and Wendun Wang. 2016. "The Forecast Combination Puzzle: A Simple Theoretical Explanation." *International Journal of Forecasting* 32 (3): 754–62. https://doi.org/10.1016/j.ijforecast.2015.12.005.

"COVID-19 Data Explorer." n.d. Our World in Data. Accessed May 30, 2021. https://ourworldindata.org/coronavirus-data-explorer.

Cramer, Estee, Evan L. Ray, Velma K. Lopez, Johannes Bracher, Andrea Brennen, Alvaro J. Castro Rivadeneira, Aaron Gerding, et al. 2021. "Evaluation of Individual and Ensemble Probabilistic Forecasts of COVID-19 Mortality in the US." *medRxiv*, February, 2021.02.03.21250974. https://doi.org/10.1101/2021.02.03.21250974.

Cramer, Estee, Nicholas G Reich, Serena Yijin Wang, Jarad Niemi, Abdul Hannan, Katie House, Youyang Gu, et al. 2020. "COVID-19 Forecast Hub: 4 December 2020 Snapshot." Zenodo. https://doi.org/10.5281/zenodo.3963371.

ECDC. 2021. "Forecasts of New Cases and Deaths Due to Covid-19 over the Next Four Weeks in Countries Across Europe and the UK." https://covid19forecasthub.eu/.

Fay, Colin, Vincent Guyader, Sébastien Rochette, and Cervan Girard. 2021. *Golem: A Framework for Robust Shiny Applications*. https://github.com/ThinkR-open/golem.

Ferguson, N., D. Laydon, G. Nedjati Gilani, N. Imai, K. Ainslie, M. Baguelin, S. Bhatia, et al. 2020.

"Report 9: Impact of Non-Pharmaceutical Interventions (NPIs) to Reduce Covid19 Mortality and Healthcare Demand." Report. *20*. https://doi.org/10.25561/77482.

Funk, Sebastian, Sam Abbott, B. D. Atkins, M. Baguelin, J. K. Baillie, P. Birrell, J. Blake, et al. 2020. "Short-Term Forecasts to Inform the Response to the Covid-19 Epidemic in the UK." *medRxiv*, November, 2020.11.11.20220962. https://doi.org/10.1101/2020.11.11.20220962.

Funk, Sebastian, Anton Camacho, Adam J. Kucharski, Rachel Lowe, Rosalind M. Eggo, and W. John Edmunds. 2019. "Assessing the Performance of Real-Time Epidemic Forecasts: A Case Study of Ebola in the Western Area Region of Sierra Leone, 2014-15." *PLOS Computational Biology* 15 (2): e1006785. https://doi.org/10.1371/journal.pcbi.1006785.

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.

Gneiting, Tilmann, and Adrian E. Raftery. 2005. "Weather Forecasting with Ensemble Methods." *Science* 310 (5746): 248–49. https://doi.org/10.1126/science.1115255.

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.

Hamner, Ben, and Michael Frasco. 2018. *Metrics: Evaluation Metrics for Machine Learning*. https://CRAN.R-project.org/package=Metrics.

Held, Leonhard, Sebastian Meyer, and Johannes Bracher. 2017. "Probabilistic Forecasting in Infectious Disease Epidemiology: The 13th Armitage Lecture: L. HELD, S. MEYER AND J. BRACHER." *Statistics in Medicine* 36 (22): 3443–60. https://doi.org/10.1002/sim.7363.

IHME COVID-19 health service utilization forecasting team, and Christopher JL Murray. 2020. "Forecasting COVID-19 Impact on Hospital Bed-Days, ICU-Days, Ventilator-Days and Deaths by US State in the Next 4 Months." *medRxiv*. https://doi.org/10.1101/2020.03.27.20043752.

Johansson, Michael A., Karyn M. Apfeldorf, Scott Dobson, Jason Devita, Anna L. Buczak, Benjamin Baugher, Linda J. Moniz, et al. 2019. "An Open Challenge to Advance Probabilistic Forecasting for Dengue Epidemics." *Proceedings of the National Academy of Sciences* 116 (48): 24268–74. https://doi.org/10.1073/pnas.1909865116.

Jordan, Alexander, Fabian Krüger, and Sebastian Lerch. 2019. "Evaluating Probabilistic Forecasts with scoringRules." *Journal of Statistical Software* 90 (12): 1–37. https://doi.org/10.18637/jss.v090.i12.

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/.

McAndrew, Thomas, Nutcha Wattanachit, Graham C. Gibson, and Nicholas G. Reich. 2021. "Aggregating Predictions from Experts: A Review of Statistical Methods, Experiments, and Applications." *WIREs Computational Statistics* 13 (2): e1514. https://doi.org/10.1002/wics.1514.

McGowan, Craig J., Matthew Biggerstaff, Michael Johansson, Karyn M. Apfeldorf, Michal Ben-Nun, Logan Brooks, Matteo Convertino, et al. 2019. "Collaborative Efforts to Forecast Seasonal Influenza in the United States, 2015–2016." *Scientific Reports* 9 (1, 1): 683. https://doi.org/10.1038/s41598-018-36361-9.

Metaculus. 2020. "A Preliminary Look at Metaculus and Expert Forecasts." June 22, 2020. https://www.metaculus.com/news/2020/06/02/LRT/.

Tetlock, Philip E., Barbara A. Mellers, Nick Rohrbaugh, and Eva Chen. 2014. "Forecasting Tournaments: Tools for Increasing Transparency and Improving the Quality of Debate." *Current Directions in Psychological Science* 23 (4): 290–95. https://doi.org/10.1177/0963721414534257.

Viboud, Cécile, Kaiyuan Sun, Robert Gaffey, Marco Ajelli, Laura Fumanelli, Stefano Merler, Qian Zhang, Gerardo Chowell, Lone Simonsen, and Alessandro Vespignani. 2018. "The RAPIDD Ebola Forecasting Challenge: Synthesis and Lessons Learnt." *Epidemics*, The RAPIDD Ebola Forecasting Challenge, 22 (March): 13–21. https://doi.org/10.1016/j.epidem.2017.08.002.

Yamana, Teresa K., Sasikiran Kandula, and Jeffrey Shaman. 2016. "Superensemble Forecasts of Dengue Outbreaks." *Journal of The Royal Society Interface* 13 (123): 20160410. https://doi.org/10.1098/rsif.2016.0410.

Yan, Yachen. 2016. *MLmetrics: Machine Learning Evaluation Metrics*. https://CRAN.R-project.org/package=MLmetrics.

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT
United Kingdom
Switchboard: +44 (0)20 7636 8636

www.lshtm.ac.uk

Observational / Interventions Research Ethics Committee

Mr James Munday
LSHTM

9 November 2020

Dear Mr James Munday

Study Title: Forecasting of COVID-19 using human predictions

LSHTM Ethics Ref: 22290

Thank you for responding to the CaRR Committee's request for further information on the above research and submitting revised documentation.

The further information has been considered on behalf of the Committee by the Chair.

Confirmation of ethical opinion

On behalf of the Committee, I am pleased to confirm a favourable ethical opinion for the above research on the basis described in the application form, protocol and supporting documentation as revised, subject to the conditions specified below.

Conditions of the favourable opinion

Approval is dependent on local ethical approval having been received, where relevant.

Approved documents

The final list of documents reviewed and approved by the Committee is as follows:

Document Type	File Name	Date	Version
Investigator CV	joel	08/10/2020	1
Investigator CV	james	08/10/2020	1
Investigator CV	kath	08/10/2020	1
Investigator CV	nikos	08/10/2020	1
Investigator CV	sam	08/10/2020	1
Investigator CV	seb	08/10/2020	1
Investigator CV	sophie	08/10/2020	1
Protocol / Proposal	Human Forecasting Study Protocol	09/10/2020	1
Investigator CV	robin	09/10/2020	1
Other	citi_completion_report	16/10/2020	1
Protocol / Proposal	Human forecasting protocol	05/11/2020	2
Information Sheet	Consent Form	05/11/2020	1
Advertisements	Recruitment Procedures	05/11/2020	1
Covering Letter	Cover Letter Ethics	05/11/2020	1

After ethical review

The Chief Investigator (CI) or delegate is responsible for informing the ethics committee of any subsequent changes to the application. These must be submitted to the Committee for review using an Amendment form. Amendments must not be initiated before receipt of written favourable opinion from the committee.

The CI or delegate is also required to notify the ethics committee of any protocol violations and/or Suspected Unexpected Serious Adverse Reactions (SUSARs) which occur during the project by submitting a Serious Adverse Event form.

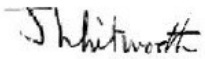
An annual report should be submitted to the committee using an Annual Report form on the anniversary of the approval of the study during the lifetime of the study.

At the end of the study, the CI or delegate must notify the committee using an End of Study form.

All aforementioned forms are available on the ethics online applications website and can only be submitted to the committee via the website at: http://leo.lshtm.ac.uk

Additional information is available at: www.lshtm.ac.uk/ethics

Yours sincerely,



Professor Jimmy Whitworth
Chair

ethics@lshtm.ac.uk
http://www.lshtm.ac.uk/ethics/

Improving health worldwide

LSHTM Ethics Application & CARE Form

Project Information

Staff members/students based at:

- (•) LSHTM
- ( ) MRC Gambia@LSHTM
- ( ) MRC Uganda@LSHTM

1. Full project title

Forecasting of COVID-19 using human predictions

2. Is this Project in fulfillment of a degree?

(•) Yes ( ) No

2a. Degree registered for

PhD

2b. Have you completed upgrading?

- ( ) Yes
- (•) No

2b (ii). If you have not yet completed upgrading, please state when upgrading is likely to take place, as well, detail why you are submitting to the ethics committee at this stage.

Upgrading should take place mid-2021.
We are submitting the form at this point as Covid-19 forecasts are relevant now and we are already prepared to start the project at this time.

2f(deg). Is this an original submission, or are you responding to a request for clarification from the LSHTM ethics committee?

- ( ) Original submission
- (•) Responding to request for clarification

2f(i-deg) Please upload a covering letter responding to the committee's request for clarification (please use the same format as that shown in the template cover letter available under Help-Templates). Please upload all amended documents in the relevant section of the form.

Documents

Type	Document Name	File Name	Version Date	Version	Size
Covering Letter	Cover Letter Ethics	Cover Letter Ethics.docx	05/11/2020	1	10.2 KB

Applicant Details

3d. Department of LSHTM Lead Investigator

Department of Infectious Disease I

Student Details

3a. Student details

Title	First Name	Surname
Mr	Nikos	Bosse

Telephone: +4917680165076

Email: nikos.bosse@lshtm.ac.uk

3c. Supervisor's name.

Sebastian Funk

3c (i). Supervisor's email address (if more than one, please only provide the email address of your main supervisor)

Email: sebastian.funk@lshtm.ac.uk

3 c(ii). Supervisor's institution

- (•) LSHTM
- ( ) MRC Gambia or Uganda
- ( ) Other

3e. Supervisor status

Confirmed

Project Type

Note: Completing the filter will enable and disable sections of the form so you may not see all questions.

4. Does the research involve primary data collection, analysis of data/samples that have already been collected, or a mix of both?

- (•) Primary
- ( ) Previously collected data/samples
- ( ) Mixed

4a. Is this research project classed as interventional or observational?

- ( ) Interventional
- (•) Observational

4a(ii). Select type of project:

Project involving mixed methodology (mix of qual/quan/lab etc.)

4c. Does the project involve extraction of data from patient records (e.g. medical, social care, service user records)? (This refers to primary data collection from records and does not include data that was previsously collected and is now being used in a secondary analysis).

- ( ) Yes
- (•) No

6. Does this project require review by the Commercialisation and Rapid Response (CaRR) ethics committee? (please see info icon for the remit of this committee)

- (•) Yes
- ( ) No

To allow the CaRR committee to triage and review the application as quickly as possible, please ensure you answer the following questions as accurately as possible:

1. Which local/regulatory approvals are required to carry out this work? (Q65-69 local approval section)
2. Where are you in the process of obtaining these local approvals? (Q65-69 local approval section)
3. Have you included in your application the report of an independent scientific review? Please note this is required for ethics approval and that this can be conducted in parallel to the ethics review. (Q47 funding section).
4. Based on the answer to Q2, when are you realistically able to go ahead with the project? (Q23 methods section)

Samples

6a. Does this research project involve the collection, or use of previously collected, human tissue samples e.g urine, stool, blood etc? (Please select yes even if the samples are not considered relevant material under the Human Tissue Act)

- ( ) Yes
- (•) No

6b. Will this project involve living animals (either laboratory, livestock or wild animals) AND/OR biological material that has been obtained from animals in the experiments planned?

- ( ) Yes
- (•) No

Fast-Track

7a. Will this project be conducted in conjunction with NHS staff, premises or any other connection to the NHS?

- ( ) Yes
- (•) No

7b. Is this application for fast-track? Note: MSc applications are not currently available for fast-track

- ( ) Yes
- (•) No

Vulnerable Groups

8(i). Does this research project involve vulnerable groups? Vulnerable groups include: children, individuals with mental disability or learning difficulties, pregnant women, prisoners etc (see information icon for full description).

- ○ Yes
- ◉ No

Security Sensitive Research Material

9. Does this research involve access to and/or storage of security sensitive research material? Please note that while some data is considered sensitive, such as HIV status, it is not necessarily considered security sensitive. If you are using data that could be considered sensitive, but not security sensitive please answer no to whether your research involves access to and/or storage of security sensitive research material. Please see information icon for what is considered security sensitive material.

- ○ Yes
- ◉ No

Geography

10. List the countries where the research project is to be conducted (For example: if you are conducting a secondary data analysis for your project and you will be based in the UK, select UK regardless of where the original data has come from):

United Kingdom

Please be aware that all primary health research conducted in the UK requires a sponsor. Please contact the RGIO at RGIO@lshtm.ac.uk for more information on sponsorship.

Outline

Note: Please do not copy and paste directly from the protocol. Applications where large portions of text have been copied and pasted directly from the protocol, and therefore do not properly answer the question, will be invalidated

12. Give an outline of the proposed project, including background to the proposal. Include information from any systematic reviews that have been conducted. Sufficient detail must be given to allow the Committee to make an informed decision without reference to other documents.

The primary aim of this study is to collect predictions from experts in epidemiology and infectious diseases as well as non-experts about key characteristics of the COVID-19 outbreak. To that end, experts and non-experts are asked to make short-term forecasts of Covid-19 case and death numbers using a web app. These forecasts allows to use information and resources not readily available to computer models such as knowledge about future planned policies. They also allow research into how human forecasters compare to mathematical models in terms of performance on short-term forecasts.

This study is focusing on Germany and Poland for now to make a contribution to an ongoing research project by the German Forecast Hub (https://kitmetricslab.github.io/forecasthub/forecast, https://osf.io/cy937/registrations). We were asked by the German team to contribute to the Forecast Hub and would like to use this as an opportunity to compare human forecasters against computer models. Part of that involvement has to do with the fact that two of the investigators (Sebastian Funk and Nikos Bosse) are German and know the German researchers, but the group is also contributing (computer-generated models) to other international efforts like the US Forecast Hub (https://github.com/reichlab/covid19-forecast-hub). At a later point we would also like to set up a similar human expert project for the UK.

12a. Upload the study protocol (compulsory for staff and doctoral students), including data collection forms, questionnaires and topic guides. Please upload each document separately, ensuring that the date and version number of each document is correct.

Documents

Type	Document Name	File Name	Version Date	Version	Size
Protocol / Proposal	Human Forecasting Study Protocol	Human Forecasting Study Protocol.docx	09/10/2020	1	1.8 MB
Protocol / Proposal	Human forecasting protocol	Human forecasting protocol.docx	05/11/2020	2	2.4 MB

13. State the intended value of the project, detailing why the topic is of interest or relevance. If this project or a similar one has been done before what is the value of repeating it? Give details of overviews and/or information on the Cochrane database. This area is of increasing importance – please ensure you give a full response.

One of the challenges of the COVID-19 pandemic is the uncertainty about many of the key epidemiological parameters and, in turn, the current state and future trajectory of the outbreak. Currently, mathematical models are predominantly used to estimate the current and near-future state of the outbreak in various different countries. Mathematical models, however, are limited by the quality and availability of data and may not take into account all sources of information (e.g. recent government announcements or newly-released scientific studies). Eliciting forecasts from human forecasters is one way in which estimates may be able to incorporate newly-available information in real time.

Forecasts from human forecasters will be aggregated and submitted to the German Forecast Hub (https://github.com/KITmetricslab/covid19-forecast-hub-de) and possibly to the US Forecast Hub (https://github.com/reichlab/covid19-forecast-hub) and similar projects to help inform important policy decisions. This is a unique and novel contributions as none of the models previously registered make use of human forecasters.

The project also allows more research into how we can best improve human forecasters. We aim to show forecasters different baseline models to investigate how anchoring affects human forecasters.

15. Overall aim of project

The primary aim of this study is to collect predictions from experts in epidemiology and infectious diseases as well as non-experts about the short-term trajectory of Covid-19.

16. Specific objectives of project

The data collected through this study will serve three outcomes.

Firstly, short-term forecasts will be used to contribute forecasts to research institutions like the German Forecast Hub (https://github.com/KITmetricslab/covid19-forecast-hub-de), representing an alternative approach to existing mathematical models.

Secondly, we intend to use the collected forecasts to investigate how baseline models change human forecasts.

Finally, results from this study can inform whether expert elicitation can be a useful tool to support public health decision making during infectious disease outbreaks.

At a later point we would like to expand this project to expert forecasting in the UK, but this is not currently the main objective.

Methods

Note: Please do not copy and paste directly from the protocol. Applications where large portions of text have been copied and pasted directly from the protocol, and therefore do not properly answer the question, will be invalidated

18. Specify the procedures/methodology to be conducted during the project. Please include outcome measures and plans for data management and analysis. For literature reviews, include details on search strategy, search terms, inclusion and exclusion criteria.

Participants are invited via a link to complete a weekly survey using a web app. The app asks users to provide a predictive distribution for Covid-19 case and death numbers in multiple locations (Currently planned are Germany and Poland).

We expect experts to make up a large fraction of respondents, but really everyone who has an interest in joining this effort is invited to participate. Given this project will be approved we intend to do the following (or similar things) to invite participants
- invite members of the CMMID through Slack to participate
- circulate this among other teams contributing to the German (and possibly) US Forecast Hub
- Sending a link to the app to other research groups like the Newton Institute that might be interested
- inviting friends who may be interested to join
- making a post on Twitter / Reddit / facebook / on blogs about the project and invite participation

We think of this process as a rather informal invite of anyone who may be interested to join, so we may not tell in advance who exactly will participate.

In order to use the app, users must create a simple user account. The minimum information needed is a username and a password. Participants are encouraged to share their name as well as their e-mail address so that we can contact them with questions and send a weekly reminder to complete the survey. This, however, is not required and participants are also allowed to provide a fake name and no e-mail address. If they wish, participants can have their username appear on a performance board to obtain feedback on their performance. Weekly reminders will be sent as long as the project continues or if requested otherwise by a participant.

Submitted responses will be saved in a Google Sheet stored in a Google Drive folder that only the aforementioned investigators have access to. Responses will be stored in two separate sheets: The sheet with that holds the forecasts will not include the forecasters name, but instead only a random unique forecaster ID. A second sheet will hold the personal information (name, email address, institutional affiliation) as well as the encrypted password and the random unique forecaster ID. Whenever we process the data in any way, the random forecaster ID will be used to respect the user's privacy. Forecasts will be removed from Google Drive once a week and stored on LSHTM hard drives. Pseudonymised forecasts may be made available to others through sites like Github. We will not share participants' personal data under any circumstances. Participants will be asked for their explicit consent when they create a user account and can have their data deleted upon contacting us.

Data management will follow the LSHTM Confidentiality and Anonymisation of Research Data SOP (LSHTM-SOP-036-01). The data will be stored in a secure drive on with access held only by the investigators. Individual-level data will be pseudonymised and identified with a unique participant ID; information linking participants' ID number to their identity (names, emails, institution, or other identifying information) will be stored separately and will remain private.

Processing and analysis of the pseudonymised data will be performed using the statistical programming language R. Individual estimates will be pseudonymised and identified by a unique participant ID. Forecasts will be aggregated using different methods, e.g. a median ensemble or a perfomance-weighted ensemble. Different aggregation techniques will be tested to learn more about their performance. Aggregated forecasts will be submitted to the German Forecast Hub (https://github.com/KITmetricslab/covid19-forecast-hub-de) and possibly similar institutions.

Forecasts will be scored and evaluated using metrics like the weighted interval score once ground truth data becomes available. These metrics are already implemented in R packages developed by the research team. Performance will be published on a leader board, where those participants who provided a name to appear on the leader board can learn about their performance.

20. Please specify the total number of participants to be recruited into the research project.

We aim to enroll a minimum of 20 participants initially, but will continue to recruit new experts throughout the process.

20a. Please provide the scientific justification for the sample size. Please include justification for the age, gender, source and method of recruiting participants for the research project.

The minimum number of experts required for a robust expert elicitation procedure is generally cited as being between 5 and 20, but we should also account for the fact that participants may not complete the study every week, as well as attrition of participants over the course of the study.

23. Proposed start date of the project

24/10/2020

24. Proposed end date of the project

31/10/2021

Risks and Discomforts

29. Give details of all clinical and non-clinical procedure(s) that will be received by participants as part of the research protocol. Non-clinical procedure(s) can include seeking consent, interviews, non-clinical observations and use of questionnaires. Clinical procedure(s) can include uses of medicinal products or devices, other medical treatments or assessments, mental health interventions, imaging investigations and taking samples of human biological material. Include procedures which might be received as routine clinical care outside of the research.

none

29a. Please provide details of who will conduct the procedure, the average time taken per procedure (minutes, hours or days) and where it will take place.

The study will be conducted via an online survey. The study will be longitudinal, with participants being asked to complete one survey per week, and each survey will take approximately 5-10 minutes to complete. The survey will be conducted using a R shiny web app.

29b. State the potential discomfort, distress or hazards that research participants may be exposed to (these may be physical, biological and/or psychological) as a result of all procedure(s).

None

29c. What precautions are being taken to control and modify these? Include information on hazardous substances that will be used or produced, and the steps being taken to reduce risks.

NA

Experience

30. State the personal experience of the applicant and of senior collaborators in the research project in the field concerned, and their contribution to this project. Indicate any previous work done related to the project topic including student and/or professional work, or publications

All members of the research team have experience in the field of infectious disease epidemiology and mathematical modelling of infectious diseases. The team is working with guidance from collaborators in the United States of America who are conducting similar work on influenza-like diseases in the USA.

30a. Upload the CVs for all main investigators working on the project. For MSc students, please upload your CV only.

Documents

Type	Document Name	File Name	Version Date	Version	Size
Investigator CV	joel	joel.docx	08/10/2020	1	15.7 KB
Investigator CV	james	james.doc	08/10/2020	1	70.0 KB
Investigator CV	kath	kath.pdf	08/10/2020	1	138.6 KB
Investigator CV	nikos	nikos.pdf	08/10/2020	1	69.9 KB
Investigator CV	sam	sam.pdf	08/10/2020	1	53.7 KB
Investigator CV	seb	seb.pdf	08/10/2020	1	79.6 KB
Investigator CV	sophie	sophie.pdf	08/10/2020	1	52.4 KB
Investigator CV	robin	robin.pdf	09/10/2020	1	252.4 KB

30e. Have the main investigators undertaken any Research Ethics/Human Subjects Protection training (either online or face to face)? (links to online training can be found in the information icon)

- (•) Yes
- ( ) No

30e(i). Please upload a copy of the certificate(s) (if available)

Documents

Type	Document Name	File Name	Version Date	Version	Size
Other	citi_completion_report	citi_completion_report.pdf	16/10/2020	1	129.2 KB

Informed Consent - Primary

If any photographs are to be taken, whether for teaching or research purposes, ensure that the participant's consent to their use has been given in line with the provisions in British Medical Journal, 1998, 316, 1009-1011.

32. Who will be responsible for taking consent and what training/experience do they have?

Nikos Bosse (nikos.bosse@lshtm.ac.uk) is responsible for the data collection and designed the consent form. Nikos has received training on the ethics of conducting scientific research in a course taken at the Universiy of Göttingen, Germany.

32a. Will you be obtaining written consent?

- (•) Yes
- ( ) No

32a(i). State the manner in which consent will be obtained (how and from whom). Where appropriate, state how the information and consent form will be translated into local languages.

Users will be asked to consent to the following form when they sign up by pressing a button "I understand and consent".

32b. Do you expect any of your potential participants to be illiterate?

- ( ) Yes
- (•) No

32f. Please upload the information sheet(s) and consent form(s). Please upload each document separately, ensuring that the date and version number of each document is correct.

Documents

Type	Document Name	File Name	Version Date	Version	Size
Information Sheet	Consent Form	Consent Form.docx	05/11/2020	1	9.9 KB

32g. Upload recruitment procedures (eg advertisements, emails, posters). Please upload each document separately, ensuring that the date and version number of each document is correct.

Documents

Type	Document Name	File Name	Version Date	Version	Size
Advertisements	Recruitment Procedures	Recruitment Procedures.docx	05/11/2020	1	7.1 KB

Payments

37. Will payments be made to participants? These should usually not be for more than travelling expenses and/or loss of earnings and must not represent an inducement to take part.

- ( ) Yes
- (•) No

Confidentiality & Data

39. Specify how confidentiality will be maintained with respect to the data collected. When small numbers are involved, indicate how possible identification of individuals will be avoided. Where data will be anonymised, specify how this will be done.

Individual-level data will be pseudonymised and identified with a unique random participant ID. For all outward facing publication of forecast we will only use the random participant ID to identify forecasters. Information linking participants' ID number to their identity (name and e-mail if provided by the forecaster) will be stored separately to the main data and will remain private.

We do not expect small participant numbers to be a concern in this study as participants can very easily enter. If we see that participant numbers are low we can artificially increase numbers by creating numerous toy forecasts with a set of participant IDs only known to us. This will prevent anyone from identifying participants who do not wish to be identified.

It is, in principle conceivable that someone can link a name published on a leader board to the random participant ID based on performance. We do, however, not believe that this of great concern. If a participant chose to have their username published on the leader board their identity is already somewhat public. If the participant chose to provide a fake name, then the only connection that can be made is between a fake name and a random participant ID - none of these contain any private information on individuals.

40. State how your data will be stored and what will be done with it at the end of the project.

Data management will follow the LSHTM Confidentiality and Anonymisation of Research Data SOP (LSHTM-SOP-036-01). Data that is submitted through the web app will be stored on a Google Sheet in a Google Drive folder only accessible to the investigators. Data that identifies forecasters will be stored separately from the forecasts.

Once a week forecasts will be cleared from the Google Drive folder. Pseudonymised forecasts will then be stored in a repository on github.com.

All e-mail communication (weekly reminders) will be conducted using the e-mail address "nikos.bosse@lshtm.ac.uk". Trace of the e-mail communication will be deleted after every weekly reminder. For any other inquiries and communication with participants, mails will be deleted after the purpose of that communication has been fulfilled.

All personal data (name, e-mail address) will be deleted at the end of the research project. Participants can request the deletion of all personal information at any time.

41. Are there plans to share the data, or add the data to a repository in the future?

- (•) Yes
- ( ) No

If yes, please be aware of the following:

Explicit consent should be obtained from participants regarding the possible use of their anonymised data in the public domain via a data repository.

42. How will the data be shared and what safeguards are in place to ensure the use of the data is for valid research?

The data will be shared in the interest of open science and reproducibility through github.com. Aggregated forecasts will be shared with the German Forecast Hub and possibly other Forecast Hubs.

We do not expect potential misuse of data to be of major concern.

Funding

46. Do you have external funding for this project?

- (•) Yes
- ( ) No

46a. Please provide the Letter of Intent reference number

626IDE

46a(i). Please provide the name of the funder

Wellcome Trust

46a(ii). Please include details of the funding available for this project.

Senior Research Fellowship to Sebastian Funk.

Date grant accepted or funding agreed:

06/08/2018

Date end of funding:

31/03/2020

46c. Are you in receipt of any funding from the United States? Or will you be collaborating with (or with individuals from) a US Institution/organisation?

- ( ) Yes
- (•) No

47. Has the project been sent out for peer/independent scientific review (please select yes if the project is being sent to the SCC)?

- ( ) Yes
- (•) No

47a. If no, why has the project not been sent peer/independent scientific review?

This is a rapid data collection as part of the response to COVID-19. The methodology will be subject to extensive review internally.

49. Does the Chief Investigator or any other investigator/collaborator have any direct personal involvement (e.g. financial, share holding, personal relationship etc.) in the organisations sponsoring or funding the research that may give rise to a possible conflict of interest?

- ○ Yes
- ◉ No

50. Will individual researchers receive any personal payment over and above normal salary, or any other benefits or incentives, for taking part in this research?

- ○ Yes
- ◉ No

Local Approval

66. For all countries listed in Q9, please provide details of the arrangements being made to obtain local ethical and/or regulatory approval. Please electronically append copies of local approval letter(s) where this has already been obtained. Where you believe local approval is not required, please explain why not and describe any less formal permissions, invitations or support you are being given for this work. Upload local permission letters as applicable. (Where the research is to take place overseas, ethical approval must be obtained in the country(s) concerned. Approval from the LSHTM Committee is dependent on local approval having been received. You MUST NOT start your project until all relevant approvals are in place.)

None required

66a. Where the research is taking place in the UK, please list other UK Committees (including other LSHTM ethics committees) from which approval is being, or has been, sought.

None required

DPIA Screening

132. Does your research involve the processing of identifiable data in the UK at any stage? (for the purposes of this question pseudo-anonymised data is still considered identifiable. Please see the information icon for additional information on the anonymisation of data.)

- ◉ Yes
- ○ No

133. Have you already/will you be carrying out a Research Data Protection Impact Assessment for your project? (please see information icon for guidance on when a DPIA is required)

- ○ Yes
- ◉ No

134. Why have you not/are you not planning to carry out a Research Data Protection Impact Assessment for your project?

- ○ A DPIA is not required
- ◉ Unsure of whether a DPIA is required

If you are unsure of whether a DPIA will be required for your project please answer the questions below. If you answer Yes to one or more of the below questions a DPIA may be required for your project and you should contact the Schools Data Protection Officer (DPO) for further advice. The DPO can be contacted at DPO@lshtm.ac.uk

136. Does your research project involve data relating to vulnerable individuals (for example, children or vulnerable adults?

- ○ Yes
- ◉ No

137. Will your research project involve the use of biometric data or genetic data?

- ○ Yes
- ◉ No

138. Will your research project involve the processing of personal data on a large scale?(please see information icon for more information)

- ○ Yes
- ◉ No

139. Will your research project involve the evaluation or scoring of personal data, including profiling and predicting, in particular relating to aspects concerning individuals' performance at work, economic situation, health, personal preferences or interests, reliability or behaviour, location or movements (for example, genetic testing)?

- ○ Yes
- ◉ No

140. Will your research project involve the systematic monitoring of individuals, including in a publicly accessible area?

- ○ Yes
- ◉ No

141. Does your research project involve the use of innovative technology or the novel application of existing technology (including artificial intelligence, machine learning, deep learning, etc)?

- ○ Yes
- ◉ No

142. Will your research project involve making decisions about, or taking actions against, individuals by automated means which might produce legal effects concerning them, or similarly significantly affect them through decisions made?

- ○ Yes
- ◉ No

143. Will your research project involve the processing of personal data which may result in preventing individuals from exercising a right or making decisions about individuals' access to a product, service, opportunity or benefit (including e.g. treatment)?

- ○ Yes
- ◉ No

144. Does your research project involve data matching i.e. combining, comparing or matching personal data obtained from multiple sources?

- ○ Yes
- ◉ No

145. During your research project, will you be processing personal data that have not been obtained direct from the individuals and the individuals concerned will be unaware that LSHTM is collecting and using their personal data?

- ○ Yes
- ◉ No

146. Does your research project involve tracking individuals' geolocations or behaviour, including but not limited to, the online environment?

- ○ Yes
- ◉ No

147. Does your research project involve the processing of personal data which is of such a nature that a personal data breach could jeopardise the physical health or safety of individuals?

- ○ Yes
- ◉ No

Reminder: If you have answered Yes to one or more of the above questions you should contact the DPO at DPO@lshtm.ac.uk for further advice regarding whether a DPIA is required for your project.

Signature Instructions

The form should be completed and finalised prior to signing or requesting signatures. Students should ensure that the Supervisor signs prior to the Course Director/Project Module Organiser. For external supervisors, please ensure that they have registered for an account prior to requesting the signature.

Signature - Applicant

Student Declaration

- ☑ I have read and understood, and agree to abide by the LSHTM Good Research Practice policy as well as all applicable Standard Operating Procedures, including on informed consent
- ☑ I undertake to abide by all regulations, guidelines and standards of good practice, including but not limited to the Data Protection Act 2018, GDPR, and the Declaration of Helsinki
- ☑ I undertake to abide by all local rules/laws for non-UK research
- ☑ I agree to conduct my project on the basis set out in this form, and to consult staff (initially, my Supervisor) if making any subsequent changes
- ☑ I agree to inform the ethics committee of any changes made to this form, and will not implement any changes until approval from the ethics committee has been received
- ☑ I undertake to adhere to all conditions set out by review bodies in giving approval and will not start the project until all required approvals are in place
- ☑ I agree to comply with the relevant safety requirements, and will submit a separate request for LSHTM travel insurance where relevant
- ☑ I agree to inform the Faculty Safety Officer and/or the Off-Site Safety Advisor (as required) if there are any changes to the risk assessment
- ☑ I confirm that there are no conflicts of interest that preclude my participation in the project

Signed: This form was signed by Mr Nikos Bosse (nikos.bosse@lshtm.ac.uk) on 05/11/2020 8:44 PM

Signature - Supervisor

Supervisor signature

I declare that:

- I agree that the information submitted in this application is a reasonable summary of the proposed project.
- I agree that this form correctly indicates whether or not ethics approval will be required.
- I agree that this form contains adequate information for the ethics committee to form an opinion of the proposed project.
- I agree that all required supporting documentation is attached to this application.
- (For MSc projects only) I agree that responses in the Risk Assessment section address the main risks connected with a project of this nature
- I have reviewed the risk of the project, including travel, and agree that it is an acceptable risk to the student
- I confirm that there are no conflicts of interest that preclude my role as supervisor for this project
- I Have read and understood, and agree to abide by the LSHTM Good Research Practice policy

Signed: This form was signed by Dr. Sebastian Funk (sebastian.funk@lshtm.ac.uk) on 05/11/2020 9:20 PM

Signature - Other

Note:

The form will automatically submit upon receipt of all required signatures.

After submission, you will receive a confirmation email with further details.

If you have not received a confirmation email within 5 working days please email ethics@lshtm.ac.uk (staff) or MScethics@lshtm.ac.uk (students) to check the status of your submission.

This is to certify that

Nikos Bosse

successfully completed the

Research Ethics

e-learning course

with a score of

80.00 %

Comprising of modules covering:

- Introduction to the History of Research Ethics
- Fundamental Ethical Principles, including:
    - Respect for persons
    - Beneficence
    - Justice
- Responsibilities of Research Ethics Committees
- Understanding Vulnerability
- Privacy and Confidentiality

On

June 22, 2021

Provided by

London School of Hygiene & Tropical Medicine

This course meets the requirements for protection of human subjects training required by individuals involved in the design and/or conduct of National Institutes of Health (NIH) funded human subjects research.

Data Management Plan for Research Students

Project title	Exploring the role of human insight in infecctious disease forecasting
Author name	**Nikos Bosse**
Supervisor	**Sebastian Funk**
Contact email	**Nikos.bosse@lshtm.ac.uk**
Date of last edit	**11.07.2021**

Guidance on writing a Data Management Plan can be found at https://lshtm.sharepoint.com/Research/Research-data-management/ and http://servicedesk.lshtm.ac.uk

Advice and feedback can be obtained from: researchdatamanagement@lshtm.ac.uk

DESCRIBE YOUR RESEARCH

1. What digital resources – data, code, collection tools, etc. - will you collect/obtain and use?

Relevant details to mention: topics covered, type (e.g. survey), source (collected by self or others), format (e.g. STATA) and amount (e.g. 10 interviews). Draw attention to human or other data that require additional protection.

1. Collection of forecasts for COVID-19 through a self-developed shiny app

2.

3.

4.

2. What hardware and software will be used in your research?

List any hardware and software to be used, their intended purpose (e.g. collection, analysis), and (if relevant) the number needed. E.g. 20 Samsung 10" tablets, LSHTM's Open Data Kit software, STATA and MS Access for analysis.

R shiny (data collection), R (data analysis), Google Sheets (data storgage)

3. What data-related activities will be performed during the research?

List key data-related activities that you and/or others will perform during the research. For instance, trial draft survey in month 6, collect data in month 8-10, clean and anonymise data in month 11, analyse data in month 12-18.

Task	**Description**
Collect predictions	Voluntary participants make predictions of COVID-19 through a shiny app.
Aggregate predictions	Predictions will be aggregated
Submit forecasts to Forecast Hubs	Aggregated forecasts will be submitted to the German and Polish as well as the European Forecast Hub

4. What quality checks will you perform to ensure resources are fit for purpose?

Outline any quality checks to be performed before, during and after the above activities, e.g. to ensure data are captured correctly, remain accurate and complete, or ensure you avoid recognised problems. The UK Data Services offers guidance at http://ukdataservice.ac.uk/manage-data/format/quality.aspx.

Manual checks for plausibility and visual inspection.

5. How will you address ethical & legal issues within your research?

- What permissions are needed? E.g. to collect data in country, analyse data for specific purpose, share data
- From whom must approval be obtained? E.g. study participant, ethics committees, data provider
- How will permissions be provided? E.g. ask participants to sign a consent form, sign a Data Transfer Agreement

Informed consent of participants is given prior to collecting any data (by clicking ‘agree’).

Ethics approval from LSHTM is obtained.

6. What documentation will be created to ensure resources can be understood?

What aspects of the research will be documented and how? E.g. processes could be documented in Standard Operating Procedures, workflows applied described in a lab book, a codebook written to describe variables, etc.

Detailed explanation of the project is provided for the participants before any forecasts are submitted.

All code is public and documented.

STORAGE AND SECURITY

7. Where will resources be stored at key stages of your research?

Identify where resources will be held during capture, processing, analysis and other stages, and who will have access to them. Consult https://lshtm.sharepoint.com/Services/IT-Services/ServiceDesk/LSHTM-data-storage-options.pdf

Intermediate data and participant information will be stored on Google Drive. Pseudonymised forecasts are made publicly available on github.

8. What labelling conventions will you apply to manage your resources?

Briefly describe any naming conventions or classification systems you will apply to resources. E.g.

- Filenames: key characteristics you will record to group files, e.g. FG1_transcript_2018-10-01
- Variable: conventions to be used for question IDs, completed responses & missing variables
- Versions: how will you identify changes to resources over time (e.g. v1.1, v1.2)

Forecasts are stored in dated files. Version control (git) ensures changes can be tracked over time.

9. How will you keep data safe and secure? (choose one or more)

Only anonymised data will be used - personal, sensitive, or otherwise confidential data is not needed for the research		Store personal details in a separate secure location & link it via an identifier	X	Delete personal & confidential details at earliest opportunity (specify when below)	X
Use digital storage that require a username/password or other security feature	X	Physical security (such as locked cabinet or room)		Protect portable devices using security features, e.g. biometric	
Encrypt storage devices		Encrypt during transfer	X	Avoid cloud services located outside EU	
Take 'Information Security Awareness training'	X	Ensure backups are also held securely			
Notes:	Any user data will be deleted as soon as the project is finished				
Identify additional steps you will take to avoid, reduce, or eliminate risks that may affect your resources.					

ARCHIVING & SHARING

10. What resources should be kept as evidence of your research?

Research often has value beyond the lifespan of the project that produced it. For this reason, many researchers are required to keep data for a set time period, typically 10 years following completion, to comply with funding or journal publication requirement. List the resources in Q1 that will be kept and for how long. If some resources can't be retained for some reason (e.g. it contains personal data), state the reason that this is not permitted.

Pseudonymised forecasts will be kept after the project has finished.

11. Where will these resources be hosted?

Identify where each resource will be hosted following research completion. E.g.

- Files intended for sharing may be hosted in the LSHTM data repository (http://datacompass.lshtm.ac.uk) or a 3rd party repository, such as UK Data Service, ArrayExpress, Zenodo, etc.
- Internal and confidential files can be held on the LSHTM Secure Server
- My supervisor will look after them

Github.com

12. When will the resources be made available? (choose one or more)

During the research life		At the same time as findings are published in an academic journal		A set time after research end, e.g. 12 months. Specify below	
Resources already available (provide details below)	X	On completion of my thesis		Other (provide details below)	
Further information / Other					
https://github.com/epiforecasts/covid.german.forecasts https://github.com/epiforecasts/europe-covid-forecasts					

13. How will you make other researchers aware that the resources exist?

Publish a metadata record describing the resources in a repository or other catalogue		Obtain a Digital Object Identifier (DOI) or other permanent ID	X
Cite resources in future research papers, e.g. in the data access statement or		Cite resources in project reports	

reference list			
Publish a description for the project website		Write and publish a Data Paper	
Add resources to a list of your academic outputs			
Other measures / Further details			

14. What steps will you take to ensure resources are easy to analyse and use in future research? (choose one or more)

Prepare a codebook or other documentation that provides an accurate description of content		Store resources in open file formats such as CSV, Rich Text, etc. See https://www.ukdataservice.ac.uk/manage-data/format/recommended-formats	X
Write a user guide that provides a high-level overview of research		Apply a standard licence that allows a broad range of uses (e.g. Creative Commons, Open Data Commons)	X
Designate a corresponding author / data custodian who will handle data-related questions		Use domain-specific standards that make it easy to import and analyse data	X
Other / Further information			

15. If resources can be made available, but not openly, what conditions on access/use must be met?

E.g. data can be used for specific types of research only. Leave blank if not applicable.

Requirement:	**To be addressed by:**

RESOURCING

16. What are the primary data management challenges in your research?

E.g. uncertainty on data management practice, data security, data-related costs, staff resources, etc.

17. How can LSHTM & others help you to better manage your data?