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 once more highlighted by the COVID-19 pandemic. Improving infectious disease forecasting therefore is an important aim. To do so, it is necessary to have a good understanding for how to evaluate predictions, as well as the necessary tools do so. While there exists an extensive literature on forecast evaluation, appropriate tools and guidelines on how to use them are under-developed. 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. Obtaining a deeper understanding of the relative strengths and weaknesses of human forecasts and model-based approaches, and how to combine them, may therefore yield important insights. Individual predictions, be it from humans or models, usually are combined into ensembles to obtain more robust and accurate forecasts. 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. In particular, it aims to improve the way how forecasts can be evaluated, and aims to obtain a deeper understanding of what role human insight should play in infectious disease forecasting and how single predictions can be aggregated to improve forecasting.

Objectives

1. Establish appropriate tools to evaluate predictions and summarise best practices in forecast evaluation. Apply these tools to to evaluate short-term forecasts of COVID-19.
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.

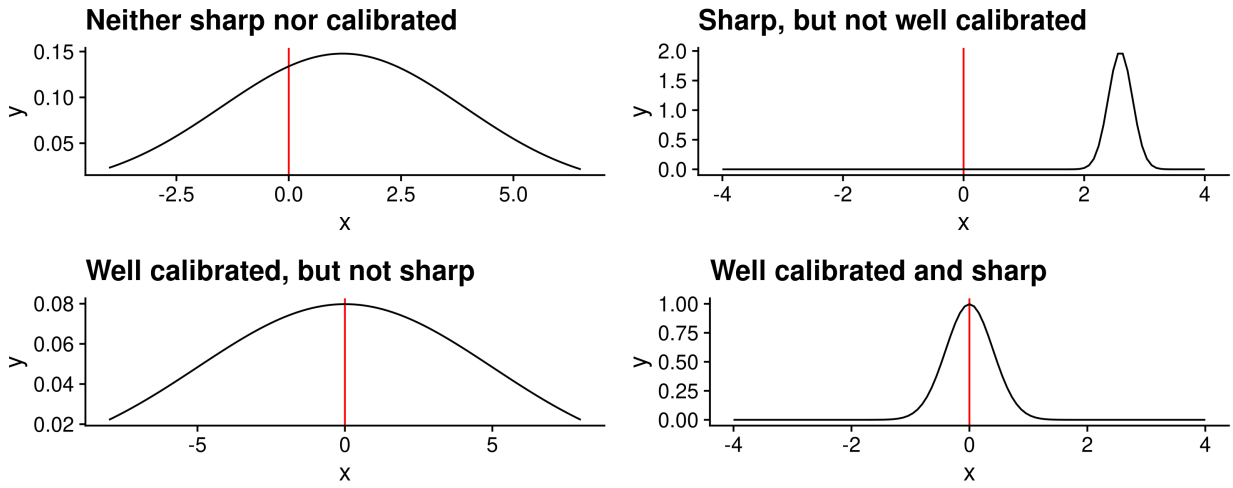


Figure 1: Schematic illustration of calibration and sharpness. True value are represented in red, the predictive distribution is shown in black.

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.

3.3.1 Evaluating calibration and sharpness independently

Evaluating calibration and sharpness independently is helpful for model diagnostics. To that end makes numerous metrics available that aim to capture different aspects of sharpness and calibration.

3.3.1.1 Assessing calibration Calibration means consistency between forecasts and observed values, but there are various ways in which a forecast can systematically deviate from the observations see [see Gneiting, Balabdaoui, and Raftery (2007) for a discussion of different forms of calibration). The `scoringutils` package allows the user to examine three different sub-aspects of calibration: bias, empirical coverage, and the probability integral transform (PIT).

Bias, i.e. systematic over- or underprediction, is a very common form of miscalibration which therefore deserves separate attention. The bias metric (with slightly different versions for the various forecast types and formats) captures a general tendency to over- and underpredict that is bound to be between minus one (underpredicton) and one (overprediction), where zero is ideal. It is derived by looking at how much of the probability mass of the predictive distribution is below or above the true observed value. For quantile forecasts we have second alternative approach available to assess over- and underprediction - by simply looking at the corresponding components of the weighted interval score. What is different between the over- and underprediction components and bias as described above is

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.

its sensitivity to outliers. The former are derived from absolute differences, while the latter is bound and rather captures a general tendency to be biased.

Another way to look at calibration (precisely: probabilistic calibration in (Gneiting, Balabdaoui, and Raftery 2007)) is to compare the proportion of observed values covered by different parts of the predictive distribution with the nominal coverage implied by the CDF of the distribution. This is most easily understood in the context of quantile forecasts, but can easily be transferred to sample-based continuous and discrete forecasts as well. To assess empirical coverage at a certain interval range, we simply measure the proportion of true observed values that fall into corresponding range of the predictive distribution. If the 0.05, 0.25, 0.75, and 0.95 quantiles are given, then 50% of the true values should fall between the 0.25 and 0.75 quantiles and 90% should fall between the 0.05 and 0.95 quantiles. We can calculate and plot these values to inspect how well different parts of the forecast distribution are calibrated. To get an even more precise picture, we can also look at the percentage of true values below every single quantile of the predictive distribution. This allows to diagnose issues in the lower and upper tails of the prediction intervals separately. A similar way to visualise the same information is a PIT histogram. In order to conveniently assess deviations between the predictive distribution and the true data-generating distribution we can transform the observed values using the probability integral transformation (PIT) (Dawid 1984). If both distributions are equal, the transformed values will follow a uniform distribution. A histogram of the transformed values can help

to diagnose systematic differences between the predictions and the observed values.

3.3.1.2 Assessing sharpness Sharpness is the ability to produce narrow forecasts. It does not depend on the actual observations and is a quality of the forecast only (Gneiting, Balabdaoui, and Raftery 2007). Sharpness is therefore only useful subject to calibration, as exemplified above in Figure 1. We may be willing to trade off a little calibration for a lot more sharpness, but usually not much. For sample-based forecasts, calculates sharpness as the normalised median absolute deviation about the median (MADN) (Funk et al. 2019). For quantile forecasts, we take the sharpness component of the WIS which corresponds to a weighted average of the individual interval widths.

3.3.2 Proper scoring rules

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.

3.3.3 Pairwise comparisons

If what we care about is to determine which model performs best, pairwise comparisons between models are a suitable approach (Cramer et al. 2021). In turn, each pair of models is evaluated based on the targets that both models have predicted. The mean score by one model is divided by the mean score of the other model to obtain the mean score ratio (see Table 2, a measure of relative performance. To obtain an overall relative skill score for a model, we take the geometric mean of all mean score ratios that involve that model (omitting comparisons where there is no overlapping set of forecasts). This gives us an indicator of performance relative to all other models. The orientation depends on the score used. For the proper scoring rules described above, smaller is better and a relative skill score smaller than 1 indicates that a model is performing better than the average model. We can obtain a scaled relative skill score by dividing a model's relative skill by the relative skill of a baseline model. A scaled relative skill smaller than one then means that the model in question performed better than the baseline.

3.3.4 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).

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 recruited mostly within the Centre of Mathematical Modeling of Infectious Diseases at the London School of Hygiene & Tropical Medicine, but participants were also invited personally or via social media to submit predictions.

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. We required that participants submitted forecasts with uncertainty that increased over time. 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 though the availability of the data evolved over the study period.

Forecasts were stored in a Google Sheet and downloaded, cleaned and processed every week for submission. If a forecaster had submitted multiple predictions for a single target, only the latest submission was kept. Some personal information (like the exact time of the forecast) was removed. 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. Inclusion was decided based on the authors' ad-hoc assessment of the validity of the forecast submission.

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

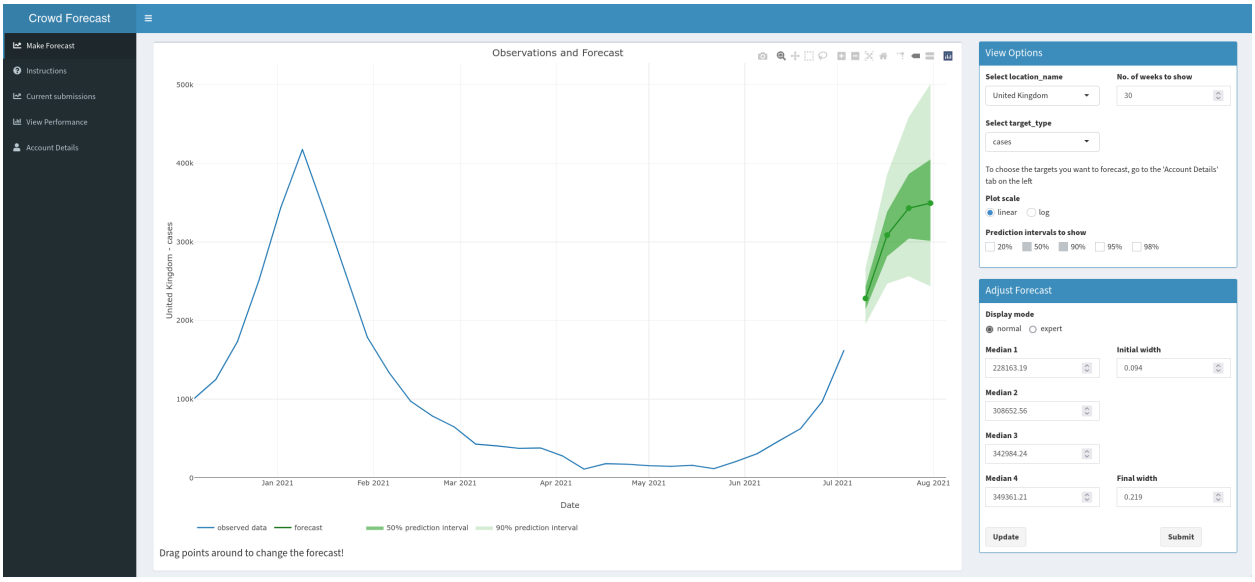


Figure 3: Screenshot of the crowdforecastr interface for the purpose of this PhD.

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.

All forecasts were processed in a Docker (Merkel 2014) container that ran automated cron jobs to ensure a reproducible environment. All code and tools necessary to generate the forecasts and make a forecast submission are available in the `covid.german.forecasts` R package (N. Bosse et al. 2020). The corresponding github repository also contains all submitted forecasts.

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. The WIS is a proper scoring rule used to evaluate forecasts in a quantile format, with lower scores representing better performance. For a growing set of equally spaced quantiles it converges to the continuous ranked probability score (CRPS) (Gneiting and Raftery 2007) that can be understood as a generalisation of the absolute error to probabilistic forecasts. The WIS can be decomposed into three separate penalties for (lack of) sharpness, overprediction and underprediction. To capture not only the absolute amount of overprediction and underprediction, we also employ a bias metric that is bound between -1 (complete underprediction, all quantiles of the predictive distribution are below the observed value) and 1 (complete overprediction, all quantiles of the predictive distribution are above the observed value) that represents a general tendency to over- or underpredict. 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). Scores were computed per forecast date, target and country and aggregated using the mean, median and standard deviation.

Aggregate Scores were then quantitatively compared and the distribution of scores was visually inspected. 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). Forecast scores for other horizons were then compared to this baseline performance. As an additional analysis, we stratified the time series into three different categories for every forecast date depending on whether numbers were monotonically rising or falling over the last two weeks prior to a given forecast date. The epidemic was categorised as either 'increasing,' 'decreasing' or 'unclear' using this categorisation. Differences of less than 5% relative to the week before were treated as zero, meaning they were interpreted as consistent with either classification.

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. Finally, we considered performance in comparison to the 'offical' hub ensemble which includes all of our forecasts except for the convolution model.

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

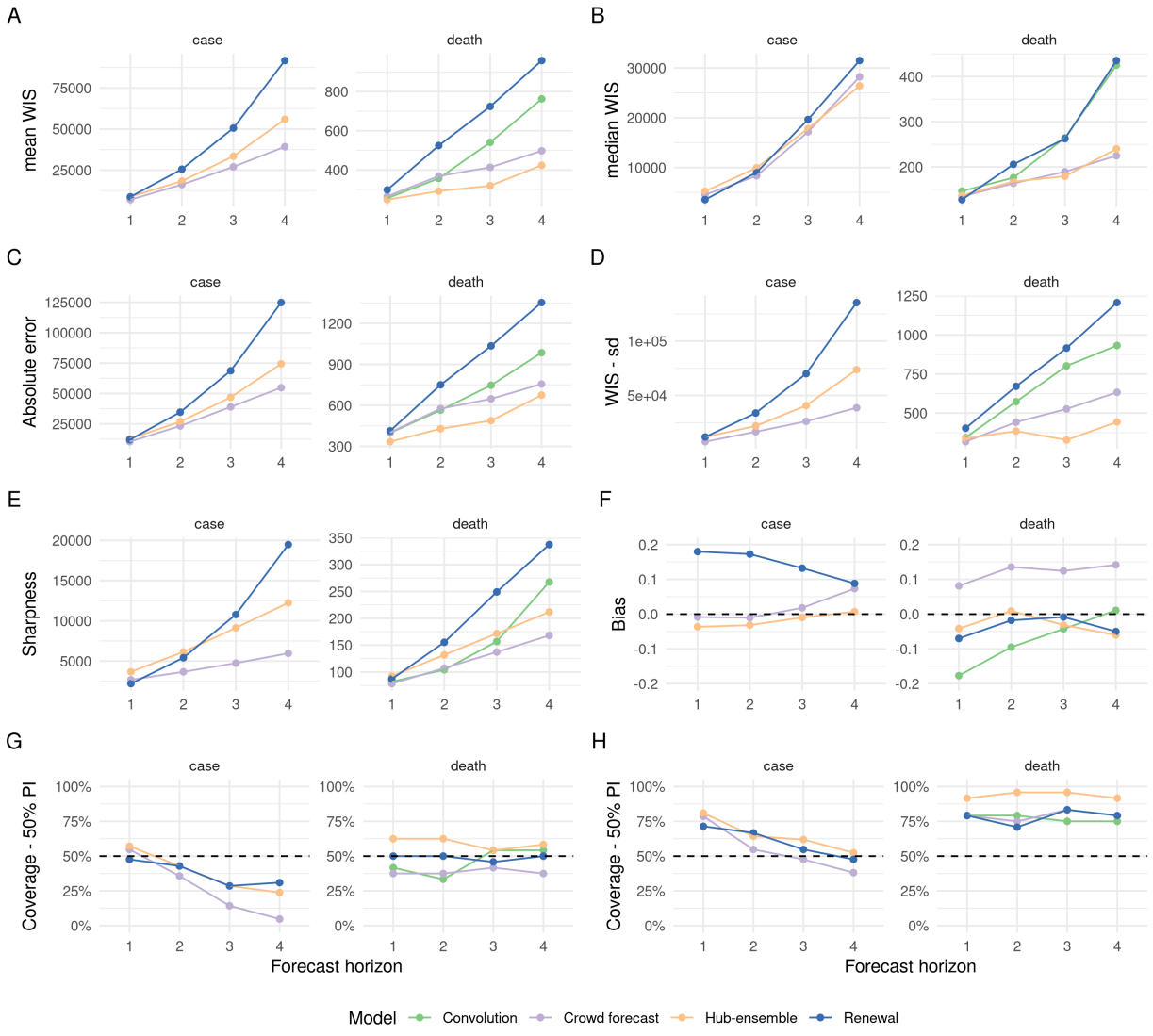


Figure 4: Visualisation of aggregate performance metrics across forecast horizons. A: mean weighted interval score (WIS) across horizons. B: median WIS. C: Absolute error of the median forecast. D: Standard deviation of the WIS. E: Sharpness (higher values mean greater dispersion of the forecast). F: 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

This work has assessed the performance of crowd-sourced human predictions and model based forecasts in a realistic real-time setting. Forecasts reflect unbiased predictive performance at the time and could not be tuned in response to reporting artifacts after submission as they were registered with an independent research organisation, timestamped and published to a public repository. The evaluation followed a methodology pre-registered by the German and Polish Forecast Hub (Bracher 2020) which makes sure the results can be fairly compared against official forecast hub evaluations. 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.

Motivating forecasters to contribute regularly proved challenging, especially given that the majority of our participants were from the UK and little connection to either Germany or Poland. In addition, lack of capacity to do proper outreach played an important role as well as a lack of time and resources to design the interface in a way that is appealing enough to attract large audiences outside of academia. Having to ask forecasters for a full predictive distribution (instead of a simple point prediction) increased complexity for participants, but allowed us submit the forecasts to the German and Polish Forecast Hub as well as analyse probabilistic aspects of human forecasts.

4.6 Conclusions

An ensemble of human insight performed as well, or better, on average than an ensemble of mathematical models. However, when evaluated in more detail, performance was mixed with the human insight ensemble performing well but not always better than other approaches. Models performed better when forecasting deaths than cases with human insight performing comparably less well, indicating that an explicit hybrid strategy may be beneficial. At longer time horizons human insight outperformed our model derived approaches. This may be partially driven by contributors implicitly accounting for further interventions. This highlights the importance of defining the role of forecasts made to inform policy as to whether or not interventions should be accounted for is a question for those consuming forecasts. The dominance of outliers on our results suggests that further work is needed to understand the importance of reliable surveillance data and the role this plays in producing good forecasts. Overall, we found that the forecasts we submitted improved ensemble performance even in instances where the individual forecasts scored poorly.

4.7 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 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 Hubs 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.

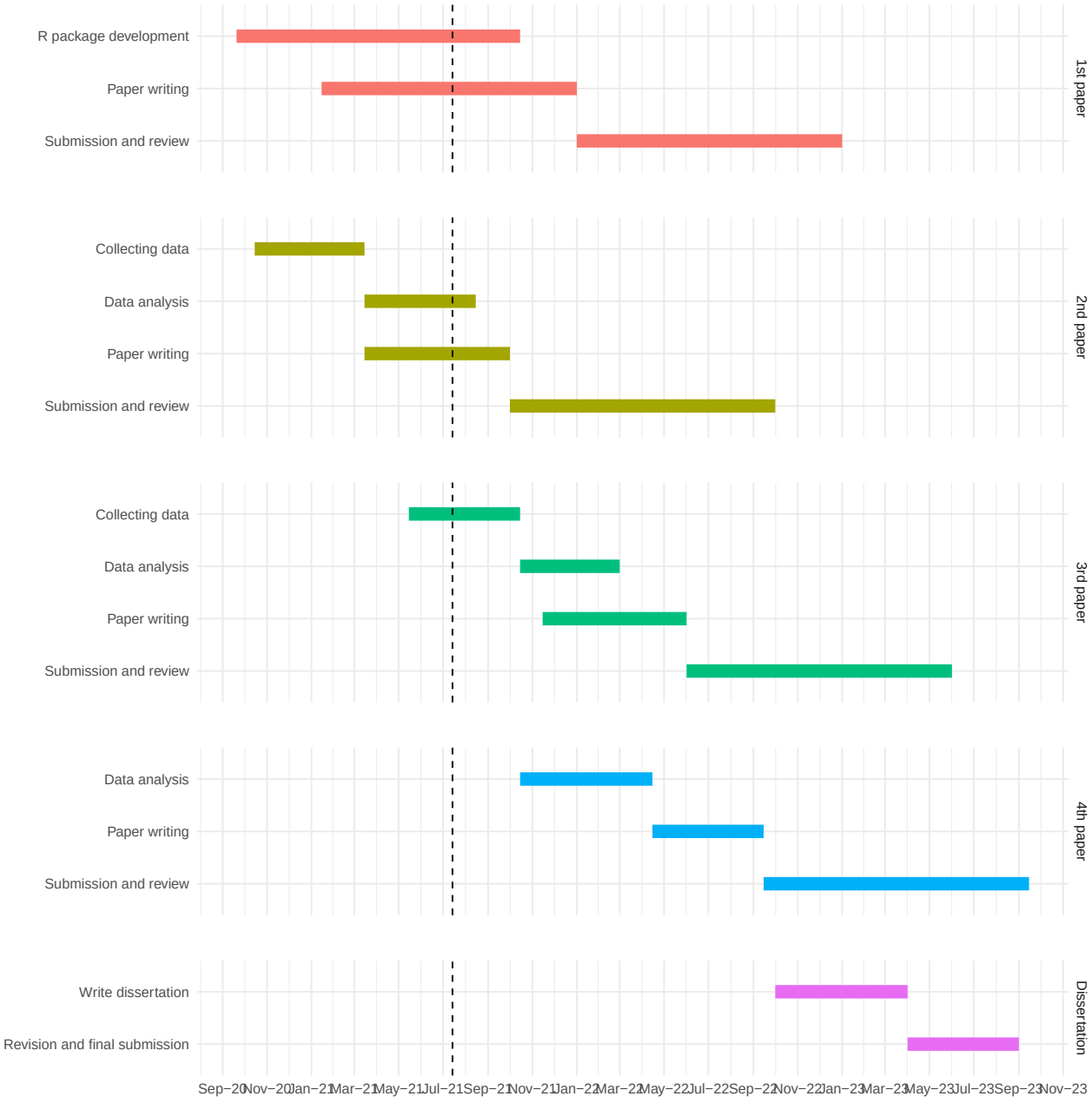
6.5 Limitations

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.

Bosse, Nikos, Sam Abbott, EpiForecasts, and Sebastian Funk. 2020. *Covid.german.forecasts: Forecasting Covid-19 Related Metrics for the German/Poland Forecast Hub*.

Bracher, Johannes. 2020. "Comparison and Combination of Real-Time Covid19 Forecasts in Germany and Poland," October. https://osf.io/k8d39.

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.

Dawid, A. P. 1984. "Present Position and Potential Developments: Some Personal Views Statistical

Theory the Prequential Approach." *Journal of the Royal Statistical Society: Series A (General)* 147 (2): 278–90. https://doi.org/10.2307/2981683.

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.

Merkel, Dirk. 2014. "Docker: Lightweight Linux Containers for Consistent Development and Deployment." *Linux Journal* 2014 (239): 2.

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.