Navigating Uncertainty - Evaluating Human and Model-Based Forecasting of COVID-19

Nikos Ioannis Bosse

June 2024

Department of Infectious Disease Epidemiology

Faculty of Epidemiology and Population Health

Centre for Mathematical Modelling of Infectious Diseases

London School of Hygiene & Tropical Medicine

Submitted in accordance with the requirements for the degree of Doctor of Philosophy of the University of London

Funded by the Health Protection Research Unit (grant code NIHR200908)

Research group affiliation: EpiForecasts

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

Junior Research Group Leader

Karlsruhe Institute of Technology

Declaration

I, Nikos Ioannis Bosse, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Nikos Bosse, June 29, 2024

Abstract

Infectious disease modelling and forecasting has garnered broad interest throughout the COVID-19 pandemic. Accurate forecasts for the trajectory of the pandemic can be useful for informing public policy and public health interventions. In this, forecast evaluation plays a crucial role. Forecasts are only useful if they are accurate. Evaluating the performance of different forecasting approaches can provide information about their trustworthiness, as well as on how to improve them. This thesis makes contributions in two areas related to forecasting and forecast evaluation in an epidemiological context. Firstly, it advances the tools available as well as our theoretical understanding of how to evaluate forecasts of infectious diseases. Secondly, it investigates the relative performance and interplay of human judgement and mathematical modelling in the context of short-term forecasts of COVID-19.

With respect to forecast evaluation, the first contribution made by this thesis is `scoringutils`, an R package that facilitates the evaluation process. The package provides a coherent framework for forecast evaluation in R and implements a selection of scoring rules, helper functions and visualisations. In particular, it supports evaluating forecasts in a quantile-based format that has recently been used by several COVID-19 Forecast Hubs in the US, Europe, and Germany and Poland. The second contribution to the field of forecast evaluation is a novel approach to evaluating forecasts in an epidemiological context. Scores like the continuous ranked probability score (CRPS) or the weighted interval score (WIS), which are common in epidemiology, represent a generalisation of the absolute error to predictive distributions. However, determining predictive performance based on the absolute distance between forecast and observation neglects the exponential nature of infectious disease processes. Transforming forecasts and observations using the natural logarithm before applying the CRPS or WIS may be more adequate in an epidemiological context. The resulting score can be understood as a probabilistic version of the relative error. It measures predictive performance in terms of the exponential growth rate and can serve as a variance-stabilising transformation assuming that the underlying disease process has a quadratic mean-variance relationship. This thesis motivates the idea of transforming forecasts before evaluating them and illustrates the behaviour of these scores using data from the European COVID-19 Forecast Hub. Log-transforming forecasts before scoring them changed the ranking between forecasters and resulted in scores that were more evenly distributed across time and space.

With respect to the role of human judgement in infectious disease forecasting, this thesis contributes two studies that analyse and compare the predictive performance of human judgement forecasts

and model-based predictions. It starts from the understanding that computational modelling, which has been the predominant way to obtain infectious disease forecasts in the past, represents a synthesis between epidemiological and mathematical assumptions and the expertise and judgement of the researchers fine-tuning the models. Understanding the interplay between human judgement and mathematical modelling better, as well as trade-offs between the two, may help make future forecasting efforts more efficient and improve predictive accuracy. This thesis uses the newly developed forecast evaluation tools to investigate the interplay between human judgement and mathematical modelling in the context of infectious disease forecasting, specifically of COVID-19. In a first study, it elicited forecasts from researchers and laypersons and compared these human judgement forecasts against predictions from a minimally-tuned mathematical model, as well as from an ensemble of several computational models submitted to the German and Polish COVID-19 Forecast Hub. It found that human judgement forecasts generally performed on par with the ensemble of computational models, performing slightly better when predicting cases and slightly worse when predicting deaths. Adding more forecasts to the ensemble was generally advantageous, even if the model to be added performed worse than the already existing ensemble. A second study replicates the basic set-up and compared human judgement forecasts of COVID-19 in the UK, elicited as part of a public "UK Crowd Forecasting Challenge", against the ensemble of all forecasts submitted to the European COVID-19 Forecast Hub. Again, forecasts performed broadly on par with the ensemble forecasts. We did not find a strong difference between self-selected "experts" and "non-experts" in terms of predictive performance. Results should generally be interpreted carefully, due to small sample sizes and susceptibility to choices made in the evaluation process. We explored a novel way to combine human judgement and mathematical modelling by asking forecasters to predict the effective reproduction number R_t which then got mapped to case and death numbers using an epidemiological model. Due to various limitations, the initial performance with this new approach was worse than that of direct human forecasts. Nevertheless, approaches that combine human judgement and mathematical modelling are promising as they could help reduce the cognitive burden of the forecasters and increase accuracy.

Ackowldedgements

I would like to express my heartfelt gratitude to my supervisor, Sebastian Funk, and my co-supervisors, Anne Cori, and Edwin van Leeuwen, as well as my advisory committee, Sam Abbott and Johannes Bracher, for their unwavering support and guidance.

Sebastian, thank you for giving me great freedom to explore and find my own way while providing exceptional guidance and advice whenever I needed it. Thank you for your friendship, and for all your kindness, patience and support.

Anne, your kind and persistent encouragement, as well as your clear and pointed suggestions have been invaluable. Thank you for being just the right amount of impatient and pragmatic and for making sure things got done in the kindest way possible.

Edwin, thank you for your warmth, your humour, your pointed questions, and insights, and all your support. It was a true pleasure to work with you and have you as a supervisor.

Sam, you were one of the main reasons I started this PhD. Thank you for enduring my occasional less-than-perfect PR and for teaching me how to do better. Thank you for all the countless hours you have spent both on scoringutils and on relentlessly supporting me with my other projects. Acts speak louder than words, and your acts have spoken very loudly. Thank you for the truly exceptional mentorship and friendship you have shown me.

Johannes, your insights on forecast evaluation have been invaluable and have made this work much better than it would have been without you. Thank you for the countless discussions on scoring rules, for your kindness, and for being always there to ensure that all things go their proper way.

Thank you to my friends and family who have supported me throughout this time.

Acronyms

Abbreviation	Meaning
CDC	Centers for Disease Control and Prevention
CDF	Cumulative density function
CFR	Case fatality rate
CRPS	Continuous ranked probability score
MCMC	Markov chain Monte Carlo
PDF	Probability density function
WIS	Weighted interval score

Contents

1 Introduction

1.1 Motivation

Much of the work presented in this PhD was motivated by issues and questions that arose from the rapid real-time response to the COVID-19 pandemic. At the time, researchers in the UK and all across the world strived to provide accurate and timely forecasts of relevant COVID-19 metrics to decision makers in a setting characterised by high uncertainty. There was uncertainty about relevant disease parameters such as the generation interval (Wallinga and Lipsitch, 2006) or transmission routes, but also uncertainty about what kinds of models were suitable to generate accurate forecasts, or how those forecasts should best be evaluated. Before COVID-19, modelling and forecasting of infectious diseases has helped inform decision making in many different settings and for various diseases such as dengue fever (Johansson et al., 2019; Yamana et al., 2016), influenza (Biggerstaff et al., 2016; Reich et al., 2019; McGowan et al., 2019). During the COVID-19 epidemic however, the attention given to infectious disease modelling and forecasting by both decision makers and the general public increased dramatically (Funk et al., 2020; Cramer et al., 2022; Bracher et al., 2021b; Sherratt et al., 2022).

The usefulness of a model or forecast, of course, depends on how accurately it can capture existing and future disease dynamics. Accuracy needs to be measured to be able to improve on existing models and forecasts or to make a decision about the degree to which they should influence decision making. When evaluating and comparing multiple models or forecasters, we can select from a large variety of scores and metrics that assess predictive performance by comparing forecasts against observed data. Different metrics reward or penalise certain behaviours of forecasts differently (Gneiting and Raftery, 2007) and the choice of the metric hence influences the result of the evaluation. It is therefore important to identify and use metrics that capture what forecast consumers actually care about (Bracher et al., 2021a; Bosse et al., 2023a). There exists an extensive body of literature on scoring rules and ways to evaluate forecasts in a variety of settings. Past evaluations of epidemiological forecasts drew from this literature, but comparatively little work went into how to score forecasts in an epidemiological context specifically (an interesting recent example is Gerding et al. (2024)). The desire for better evaluation tools and the open questions surrounding the evaluation of forecasts in an epidemiological context were the motivation for a major part of this thesis.

One aspect makes interpreting model performance and decision making in the beginning of a new disease outbreak especially challenging: data on the past accuracy of a forecaster or model is usually sparse as there are only few data points available. It is therefore important to establish a general understanding of the underlying characteristics of different types of modelling and forecasting approaches that could help estimate a priori how trustworthy different predictions may be. This might provide information on which kinds of forecasting approaches may be useful in future infectious disease outbreaks. In the context of an early

disease outbreak, resources need to be allocated to different types of modelling and forecasting approaches. For example, instead of spending resources on developing mathematical models, one could instead (or in addition) survey experts directly. Infectious disease modelling usually requires a significant amount of resources in terms of time and effort. The resulting models represent a mixture of mathematical model assumptions and human judgement required to develop and tune the model. It is therefore useful to ask how well human judgement and mathematical modelling perform in comparison and what mathematical modelling is able to add above human judgement alone. This question inspired the second major part of the work presented in this thesis.

1.2 Aims and objectives

This thesis aims to help improve the usefulness of infectious disease forecasting for public health decision making in future outbreaks such as the COVID-19 pandemic by obtaining a deeper understanding of the following.

- What is a 'good' forecast in an epidemiological context, and how we can evaluate the quality of a forecast?
- How do human judgement and mathematical modelling compare in terms of predictive performance? How do human judgment and mathematical modelling interplay and how can they best be used and combined to obtain useful forecasts?

It strives to accomplish this by fulfilling the following objectives:

- Establish appropriate tools to evaluate predictions in `R` following best practices in forecast evaluation (Paper 1, see Chapter 3).
- Develop tools to elicit human forecasts of infectious diseases, specifically COVID-19 (Paper 2, see Chapter 4)
- Analyse the role of human judgement in forecasting COVID-19 in Germany and Poland. Compare human judgement forecasts against model-based predictions and analyse the added benefit of human input over mathematical and statistical modelling (Paper 2, see Chapter 4).
- Improve current evaluation methods so that they are better suited for evaluating forecasts in an epidemiological context (Paper 3, see Chapter 5).
- Examine the potential to use public crowd forecasting tournaments to predict COVID-19 in the UK and explore possibilities to combine human judgement and epidemiological modelling (Paper 4, see Chapter 6).
- Examine how well results from one human judgement forecasting effort replicate in a different setting with a different crowd of forecasters (Paper 4, see Chapter 6).

1.3 Thesis outline

Chapter 2 provides some background for the following Chapters. It defines important terminology related to forecasting and modelling of infectious diseases used throughout this

thesis. It reviews core concepts related to forecast evaluation and proper scoring rules and provides an overview of past human judgement efforts in infectious disease forecasting.

Chapter 3 (Paper 1) introduces `scoringutils`, a software package in R that implements a selection of proper scoring rules and other evaluation metrics and offers users a coherent framework for forecast evaluation in R. The `scoringutils` package provides the basis for all forecast evaluations conducted as part of this thesis.

Chapter 4 (Paper 2) applies the evaluation methods and tools described in Chapters 2 and 3 to investigate what human judgement can contribute to the task of forecasting COVID-19. It presents a study conducted in Germany and Poland where human judgement forecasts were submitted to the German and Polish COVID-19 Forecast Hub alongside mathematical models with minimal tuning. The study was motivated both by a desire to obtain a better understanding of how to create good forecasts of COVID-19, as well as the actual need to produce timely and accurate forecasts for the German and Polish Forecast Hub. The study compares human judgement forecasts elicited using a novel open source online application against predictions from two minimally-tuned mathematical models, as well as with an ensemble of model-based predictions.

Chapter 5 (Paper 3) investigates in greater detail how forecasts should be evaluated specifically in an epidemiological context. The scoring methods discussed in Chapters 2 and 3 were not explicitly developed for application in infectious disease forecasting, but rather describe general mathematical relationships that are detached from the actual context. Evaluating the forecasts we submitted to the German and Polish Forecast Hub, as described in Chapter 4, surfaced issues with the way that forecasts are currently commonly evaluated in epidemiology. Determining predictive performance based on the absolute distance between forecast and observation, as is common practice, neglects the exponential nature of infectious disease processes. It also leads to scores that are dominated by outlier forecasts, especially during periods of high incidence, and makes it hard to compare forecasts across time, locations or forecast targets. To address these issues, Chapter 5 introduces the idea of transforming forecasts and observations before applying a score in order to obtain an evaluation that is more adequate in an epidemiological context.

Chapter 6 (Paper 4) presents a follow-up study to the one presented in Chapter 4. It analyses the results of a public forecasting challenge in the UK, applying insights from Chapter 5 on how to evaluate forecasts in an epidemiological context. Forecasts were submitted to the European COVID-19 Forecast Hub and again compared to the ensemble of all forecasts submitted to the Forecast Hub, this time also taking transformations of forecasts into account for the evaluation. In addition, Chapter 6 explores a novel way to combine human judgement and mathematical modelling by asking forecasters to predict the effective reproduction number R_t which then gets mapped to cases and deaths using an epidemiological model.

Chapter 7 discusses the results and implications of the work presented in this thesis.

1.4 Code

The code for this thesis is publicly available on GitHub[1], as is the code for the scoringutils package and the accompanying Paper (Paper 1)[2], the code for the Paper on crowd forecasts in Germany and Poland (Paper 2)[3], the code for the Paper on transforming forecasts before scoring them (Paper 3)[4], and the code for the Paper on the UK Crowd Forecasting Challenge (Paper 4)[5].

[1] https://github.com/nikosbosse/phd_thesis
[2] https://github.com/epiforecasts/scoringutils
[3] https://github.com/epiforecasts/covid.german.forecasts
[4] https://github.com/epiforecasts/transformation-forecast-evaluation
[5] https://github.com/epiforecasts/uk-crowd-forecasting-challenge

2 Background

This chapter provides definitions for terms used throughout the thesis and reviews the aspects of the literature on forecast evaluation and human judgement forecasting relevant to the remainder of this thesis.

2.1 Forecasting and Modelling

A forecast, in most general terms, is a stated belief about the future (Gneiting et al., 2007) as it will occur. Such a belief can be stated in qualitative or quantitative terms. This thesis will almost exclusively focus on quantitative forecasts.

Quantitative forecasts can either be probabilistic, or they can be point forecasts. A probabilistic forecast (Held et al., 2017) is a full predictive probability distribution over multiple possible outcomes. A point forecast, on the other hand, is a single number that represents a single outcome. A probabilistic forecast incorporates uncertainty about different outcomes in a way that a point forecast cannot. Probabilistic forecasts are therefore arguably more useful for decision making (Held et al., 2017; Ramos et al., 2013) and will be the focus of this thesis. Some authors (see e.g. Farrow et al., 2017) make a distinction between 'forecast', meaning a probabilistic forecast and 'prediction', meaning a point forecast. We will use the two terms interchangeably.

The term 'model' in its broader sense generally means a simplified representation of the world that allows someone to make statements about the future based on certain inputs. A model can be the specific understanding of the world that a human forecaster has in her mind to make sense of past and current events and that allows her to make forecasts about the future. More commonly, the term model denotes a mathematical or computer model, a set of encoded rules that describe and represent the processes that govern events in reality. Mathematical models nowadays usually use computers to map observed inputs to a model output. If not otherwise stated, we use the term 'model' to describe a mathematical model which produces a forecast (rather than the mental representation of the world in a person's head). Similarly, 'model-based' predictions mean predictions generated by a computational model. Furthermore, we use the term 'modeller' to denote a person who develops, codes, or adapts a mathematical model.

Mathematical models (see e.g. Frauenthal, 1980; Kretzschmar and Wallinga, 2009), as described above, are often also referred to as 'mechanistic models'. Mechanistic models, such as compartmental models (see e.g. Shah and Mittal, 2021) or agent-based models (see e.g. Hunter et al., 2017) explicitly model the underlying infectious disease process, encoding our understanding of infectious disease dynamics. There exists a second broad category of models that are commonly used to make forecasts in epidemiology: 'statistical models'. Statistical models derive their predictive power from the statistical relationships between different observable variables. They usually do not rely on an understanding of the underlying

processes, although they often make distributional assumptions about relevant variables. For an extensive overview of statistical time series models and their use in forecasting, see e.g. Box and Jenkins (1970); West and Harrison (1997), and Hyndman, Rob J and Athanasopoulos, George (2021). In practice, the distinction between mechanistic and statistical models is not always clear-cut. Models that make use of statistical estimation while at the same time constraining parameters based on mechanistic assumptions are common in epidemiology, and are often referred to as 'semi-mechanistic' models (see e.g. Bhatt et al., 2023, for a recent example).

We describe anyone or anything that issues a forecast as 'a forecaster'. This can either be a person voicing their judgement, or it can be a computer model or algorithm that issues a forecast based on given inputs, or a combination of both. When more clarity is required, we use the terms 'human forecaster' and either 'mathematical model' or 'computational model'.

The output of an epidemiological model is not necessarily a forecast. It could also, for example, be a nowcast, or a scenario or projection. A nowcast is a description of the world as it is in the present (in the absence of definitive data). A scenario is the representation of the future as it could look like under certain scenario assumptions, whereas a projection describes the future as it could unfold if conditions stayed the same as they were in the past (Funk et al., 2020). This is in contrast to a forecast which aims to predict the future as it will occur.

Forecasts (and nowcasts) can be judged eventually by comparing them against observed data. This is more difficult, and in many instances impossible, for scenarios or projections, as they usually make statements about a world that was not observed. While scenarios are harder to evaluate, they may be more useful for decision making. Scenarios are able to show what could occur under different assumptions and different courses of action and can therefore help inform what actions should be taken. With forecasts, it is less clear how results would change under a possible course of action. By definition, a forecast has to estimate and already incorporate possible courses of action in order to make an accurate statement about the future as it will occur.

One possibility to increase the predictive accuracy of forecasts is to combine individual forecasts into an ensemble, which usually performs better than any individual forecaster (Gneiting and Raftery, 2005; Yamana et al., 2016). To denote an ensemble of forecasts made by a group of human forecasters we will sometimes use the term 'crowd ensemble' or 'crowd forecast'. Ensembles can be either equally weighted or trained, by assigning ensemble weights based on past performance of a forecaster. Past research suggests that it is very difficult to form ensembles which outperform an equally weighted ensemble (Claeskens et al., 2016), but not impossible (Brooks et al., 2018). During the COVID-19 pandemic, forecasts for different targets have been systematically collected, aggregated and evaluated by three COVID-19 Forecast Hubs in the US (Cramer et al., 2022), Germany and Poland (Bracher et al., 2021b) and Europe (Sherratt et al., 2022).

2.2 Forecast evaluation

Forecast evaluation in a narrow sense is the process of assessing how well a forecaster's predictions align with the actual observations. In a broader sense of the term, forecast evaluation would also include elements such as an assessment of the usefulness of the forecasts to the forecast consumer, or analyses that would help understand the underlying characteristics of the model or forecaster better. Forecast evaluation helps modellers and forecasters to improve future predictions, and helps decision makers decide which forecasts to take into account when making decisions.

Conceptualy, we can think of forecasting as a game between the forecaster and nature (Gneiting et al., 2007). Nature issues a probability distribution, G, the data-generating distribution. Observed values y are drawn from this data-generating distribution G. The forecaster issues a predictive distribution, F.

For any given forecasting task, the primary aim of any forecaster should be to issue a predictive distribution F that equals the (usually unknown) true data-generating distribution G (Gneiting et al., 2007). We call a forecast F an ideal forecast if it is equal to the data-generating distribution G. For an ideal forecast, we therefore have

$$F = G,$$

where F and G are cumulative distribution functions.

2.2.1 The forecasting paradigm

Since the data-generating distribution G is usually unknown, forecasts have to be evaluated based on the predictive distribution F and the observations y alone. Gneiting et al. (2007) proposed a framework for forecast evaluation that is centred around the notion that a forecaster should aim to "maximise the sharpness of the predictive distributions subject to calibration". Calibration hereby means that the forecasts are consistent with the observations and that there are no systematic deviations between the two. In their work, Gneiting et al. (2007) distinguish several different forms of calibration (specifically: probabilistic calibration, marginal calibration and exceedance calibration). Sharpness is a property that pertains solely to the forecast and is independent of the observations. It describes how informative the predictive distribution is, i.e. how concentrated the probability mass is around any potential outcome. The opposite of sharpness is dispersion, i.e. how spread out the predictive distribution is. Sharpness and calibration are illustrated in Figure 2.1.

2.2.2 Proper scoring rules

Often, the quality of a forecast is summarised into a single number using so-called proper scoring rules (Brier, 1950; Good, 1952; Gneiting and Raftery, 2007). A scoring rule $S(F, y)$ is a function of the forecast F and the observation y that returns a single numeric value. A scoring rule is said to be proper, if under G and for an ideal forecast $F = G$, there is

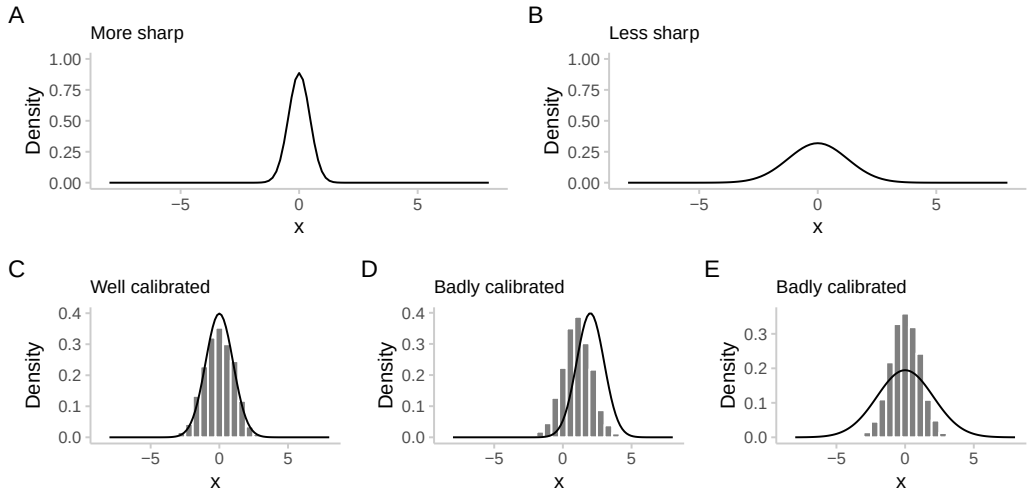

Figure 2.1: Illustration of calibration and sharpness. The probability density function of the predictive distributions is indicated by a black line, observations (draws from the unknown data-generating distribution) are represented by the grey histograms.

no forecast $F' \neq F$ that in expectation receives a better score than F. A scoring rule is considered strictly proper if, under G, no other forecast F' in expectation receives a score that is better than or the same as that of F. By convention, proper scoring rules are usually negatively oriented, meaning that smaller values are better and the best possible score is usually zero. In that sense, the score can be understood as a penalty. Among the strictly proper scoring rules most commonly used are the logarithmic scoring rule (Good, 1952) and the continuous ranked probability score (CRPS) (Epstein, 1969; Murphy, 1971; Matheson and Winkler, 1976; Gneiting and Raftery, 2007).

2.2.2.1 The logarithmic scoring rule

The logarithmic scoring rule is simply the negative logarithm of the density of the predictive distribution evaluated at the observed value

$$\text{log score} = -\log f(y),$$

where f is the predictive probability density function (PDF) and y is the observed value. For discrete forecasts, the log score can be computed as

$$\text{log score} = -\log p_y,$$

where p_y is the probability assigned to the observed outcome y by the forecast F. The logarithmic scoring rule can produce large penalties when the observed value takes on values for which $f(y)$ (or p_y) is close to zero. It is therefore considered to be sensitive to outlier forecasts. This may be desirable in some applications, but it also means that scores can easily be dominated by a few extreme values. The logarithmic scoring rule is a local scoring rule, meaning that the score only depends on the probability that was assigned to the actual

outcome (see Figure 2.2). This is often regarded as a desirable property for example in the context of Bayesian inference (Winkler et al., 1996). It implies for example, that the ranking between forecasters would be invariant under monotone transformations of the predictive distribution and the target.

2.2.2.2 The continuous ranked probability score

The continuous ranked probability score (CRPS) is popular in fields such as meteorology and epidemiology. The CRPS is defined as

$$\text{CRPS}(F, y) = \int_{-\infty}^{\infty} \left(F(x) - 1(x \geq y)\right)^2 dx,$$

where y is the observed value and F the CDF of predictive distribution. For discrete forecasts, the ranked probability score (RPS) can be used instead:

$$\text{RPS}(F, y) = \sum_{x=0}^{\infty} (F(x) - 1(x \geq y))^2.$$

The CRPS can be understood as a generalisation of the absolute error to predictive distributions (Gneiting and Raftery, 2007). It can also be understood as the integral over the Brier score (Brier, 1950) for the binary probability forecasts implied by the CDF for all possible observed values. The CRPS is also related to the Cramér-distance between two distributions and equals the special case where one of the distributions is concentrated in a single point (see e.g. Ziel (2021)). The CRPS is a global scoring rule, meaning that the entire predictive distribution is taken into account when determining the quality of the forecast (see Figure 2.2).

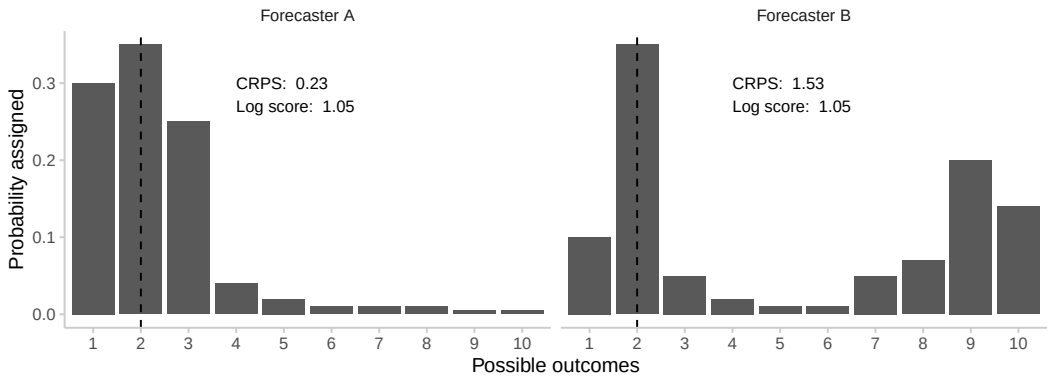

Figure 2.2: Forecasts from two forecasters, A and B, for the number of points scored by a team in a sports match (adapted from Bosse et al., 2022b). Grey bars represent the probability assigned to each outcome. The outcome later observed, 2, is marked with a black dashed line. The probability assigned to the observed outcome is the same for both forecasters. However, Forecaster A's prediction centers closely around the observed value, whereas Forecaster B allocates substantial probabilities to outcomes distant from the observed value. A local scoring rule like the logarithmic scoring rule assigns both forecasters the same score. A global scoring rule like the CRPS, which takes the full distribution into account, assigns a better score to Forecaster A.

Compared to the logarithmic scoring rule, the CRPS is also relatively more lenient when it comes to penalising poor forecasts. In particular, the logarithmic scoring rule penalises overconfidence more severely than the CRPS (see Figure 2.3A, as well as Machete, 2012). The CRPS, as a generalisation of the absolute error, grows linearly with the distance of the predictive distribution from the observed value. This is not true for the log score, which can grow quickly if the density or probability assigned to the observed outcome is small (see Figure 2.3B).

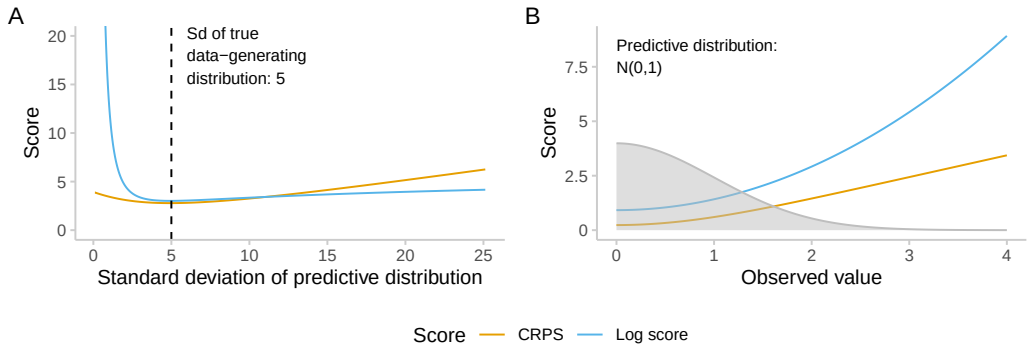

Figure 2.3: Effect of deviations from the data-generating distribution on the log score and the CRPS (adapted from Bosse et al., 2022b)). A: Effect of varying the standard deviation, while keeping the mean constant. The data-generating distribution is $\mathcal{N}(0, 5)$ (true sd is marked by the dashed black line). The standard deviation of the forecast distribution is varied along the x-axis. Coloured lines therefore represent expected scores for a data-generating distribution of $\mathcal{N}(0, 5)$ and a predictive distribution $\mathcal{N}(0, x)$. The log score penalises overconfidence (i.e. predictive distributions to the left of the dashed line, which are too sharp / underdispersed) more severely than the CRPS. B: Effect of keeping the predictive distribution constant, while varying the observed value. The predictive distribution (illustrated in grey) is $\mathcal{N}(0, 1)$. Coloured lines show the scores obtained for a standard normal forecast distribution and different observed values. The CRPS grows linearly with the observed value, while the logarithmic scoring rule produces increasingly larger penalties than the CRPS for observed values that were deemed unlikely by the predictive distribution.

Both log scores and CRPS values scale with the standard deviation of the target data-generating distribution, as forecasting an uncertain target is inherently more difficult. If the predictive distribution can be well approximated by a normal distribution, CRPS values of an ideal forecaster (for which the forecast F equals the data-generating distribution G) scale linearly with the standard distribution (Bosse et al., 2023a). This can make it difficult to compare CRPS values across forecasts for targets with differing orders of magnitudes, as will be discussed in more detail in Chapter 5. The log score mostly scales sub-linearly with the standard deviation of the data-generating distribution. This is illustrated in Figure 2.4B.

For both the CRPS and log score, a sample-based representation is available. This means that score can be estimated even in cases where the forecast is not available as a closed-form distribution, but rather is represented as a set of samples from the predictive distribution. In such cases, the CRPS has an advantage over the log score in terms of estimation. Computing the log score for continuous forecasts requires estimating a predictive density from the samples. For that reason, estimating the log score from predictive samples can often be noisy

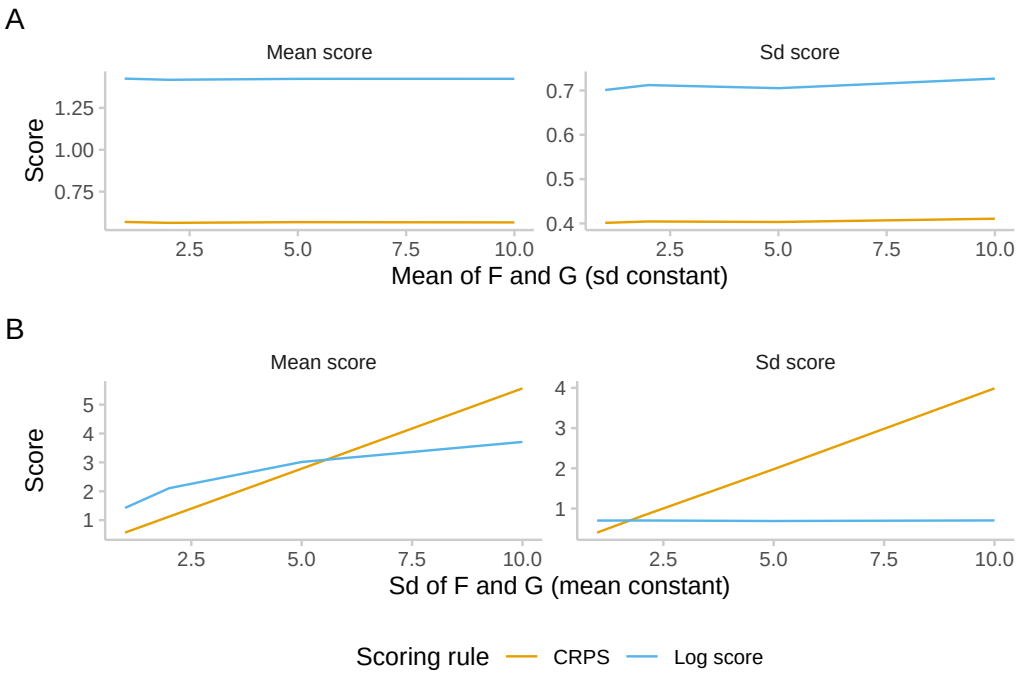

Figure 2.4: Dependency of scores on the variability of the data-generating distribution. Consider data generated from different normal distributions with differing means and standard deviations. Predictive distributions are assumed to be equal to the data-generating distribution. For every combination of mean and standard deviation, we drew 10k samples from the data-generating distribution, evaluated them, and calculated the mean and the standard deviation of the scores. A: Constant standard deviation $\sigma = 1$ of the data-generating and predictive distribution, varying mean. B: Constant mean ($\mu = 1$) of the data-generating and predictive distribution, varying standard deviation.

and require a large number of samples to work well. This is illustrated in Figure 2.5 (adapted from Jordan et al., 2019)).

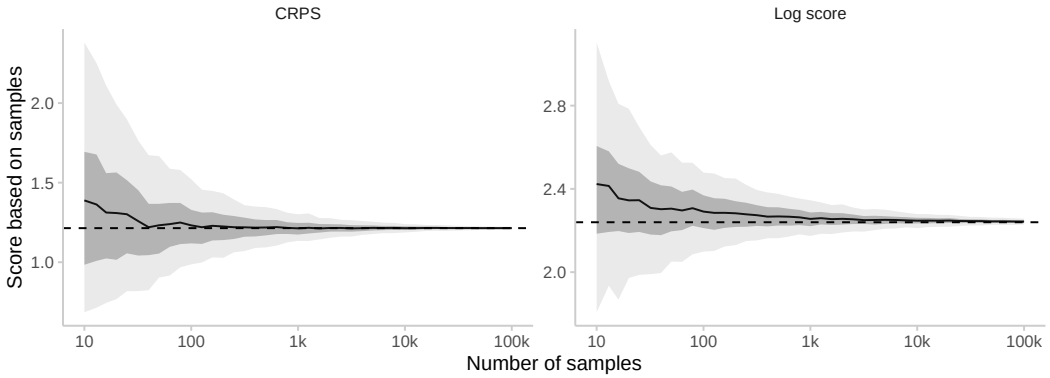

Figure 2.5: Estimates of the log score and CRPS derived from predictive samples, as adapted from Jordan et al. (2019) (see also Bosse et al. (2022b)). The observed value was assumed to be 0, the predictive distribution was $\mathcal{N}(0, 3)$. Scores were computed using samples of varying sizes from the $\mathcal{N}(2, 3)$ predictive distribution. Sample sizes ranged from 10 to 100,000. For each sample size, the process was repeated 500 times. The black line represents the mean score across these 500 repetitions, while the shaded areas indicate 50% and 90% confidence intervals. Additionally, the dashed line corresponds to the true score, calculated based on the closed-form distribution.

2.2.2.3 The weighted interval score

Recent forecasting efforts such as the COVID-19 Forecast Hubs in the US (Cramer et al., 2022), Europe (Sherratt et al., 2022), and Germany and Poland (Bracher et al., 2021b, 2022) have used a quantile format, with predictive distributions represented by a set of predictive quantiles. While predictive samples can offer richer information (for example, traces of Markov Chain Monte Carlo (MCMC) simulations can be stored that include information about spatial or temporal correlations), they are expensive in terms of storage space. Accurately representing the predictive distribution, especially in its tails, may require a large number of predictive samples. Predictive quantiles, on the other hand, require much less storage space to represent the distribution accurately (while losing information about the correlation structure between different targets).

A proper scoring rule that is well suited to evaluate forecasts in such a quantile format is the weighted interval score (WIS, see e.g. Bracher et al., 2021a; Gneiting and Raftery, 2007; Winkler, 1972, and references therein). The WIS can be understood as an approximation of the CRPS for forecasts in a quantile format. Quantiles are assumed to be the lower and upper bounds of prediction intervals symmetric around the median. The interval score for a single interval is

$$IS_\alpha(F, y) = \underbrace{(u - l)}_{\text{dispersion}} + \underbrace{\frac{2}{\alpha} \cdot (l - y) \cdot \mathbf{1}(y \leq l)}_{\text{overprediction}} + \underbrace{\frac{2}{\alpha} \cdot (y - u) \cdot \mathbf{1}(y \geq u)}_{\text{underprediction}},$$

where $\mathbf{1}()$ is the indicator function, y is the observation, and l and u are the $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$

quantiles of the predictive distribution F. l and u together form the prediction interval. The interval score can be understood as the sum of three components: dispersion, overprediction and underprediction. For a set of K prediction intervals and the median m, the score is given as a weighted sum of individual interval scores, i.e.

$$WIS = \frac{1}{K + 0.5} \cdot \left(w_0 \cdot |y - m| + \sum_{k=1}^{K} w_k \cdot IS_\alpha(F, y) \right),$$

where w_k is a weight assigned to every interval. When the weights are set to $w_k = \frac{\alpha_k}{2}$ and $w_0 = 0.5$, then the WIS converges to the CRPS for an increasing number of equally spaced quantiles (for a proof see e.g. Bracher et al., 2021a).

Scoring rules and evaluation metrics describe a mathematical relationship between forecasts and observations and capture different aspects of how a forecast deviates from observed values. This makes them objective and comparable. However, scoring rules and evaluation measures usually do not directly measure the usefulness of forecasts to forecast consumers. For example, underprediction of a target like hospitalisations could lead to more severe consequences than overprediction for a decision maker relying on a forecast in a way that's not necessarily captured by a given scoring rule. The selection of an appropriate metric therefore should ideally reflect the qualities a forecast consumer values in a forecast.

2.3 Human Judgement forecasting

Human judgement forecasting has a long history and efforts have been made in very different contexts and fields, both inside and outside of academia. Human judgement elicitation processes also differ greatly in their methodology and selection of forecasters. Efforts in the past range from surveys of experts or laypeople to structured discussions between experts aimed to produce a consensus forecast to large crowd forecasting efforts with thousands of forecasters predicting on a given question. One of the first structured methods proposed to support decision making through forecasting is the Delphi method developed by the RAND Corporation in the 1950s in the context of the Cold War (Dalkey and Helmer, 1963; Bernice Brown, 1968; Page et al., 2015). Forecasts were elicited in a structured process where experts would provide a first estimate, then discuss their estimates and potential disagreements, and then provide a final round of estimates.

Outside of academia, human predictions and estimates are routinely used to help decision making in private companies, think tanks and governments. In addition, public predictions on various topics of interest have been collected on several online prediction markets or prediction platforms. Prediction markets such as PredictIt [1], Manifold[2], Polymarket[3] and traditional betting platforms such as betfair[4] allow users to place bets on an outcome by spending either real money or token money to buy "yes" or "no"-shares in the binary outcome of a given market. Prediction platforms such as Metaculus[5], INFER[6] or Good Judgement

[1] https://www.predictit.org/
[2] https://manifold.markets/
[3] https://polymarket.com
[4] https://betfair.com
[5] https://metaculus.com
[6] https://infer-pub.com

Open[7] elicit direct forecasts from users and offer either points or sometimes monetary rewards for the most accurate users. These markets and platforms have been used in predicting a broad spectrum of events such as elections, wars, or the spread of infectious diseases.

Academic authors working on human judgement forecasting have in the past made use of such prediction platforms (McAndrew et al., 2022b,c; Tetlock et al., 2014), used more traditional means of eliciting predictions such as surveys (Recchia et al., 2021) or the Delphi method (Dalkey and Helmer, 1963; Bernice Brown, 1968; Page et al., 2015), or developed their own methods (Farrow et al., 2017). While human judgement forecasting has seen many applications in academic fields such as such as geopolitics (Atanasov et al., 2016; Tetlock et al., 2014), product forecasting (Arvan et al., 2019), or meta-science (Dreber et al., 2015; Gordon et al., 2020), its application to infectious diseases has only recently attracted more widespread attention. Notable examples of human judgement forecasting of infectious diseases include Farrow et al. (2017); Recchia et al. (2021); McAndrew and Reich (2022); McAndrew et al. (2022a,b,c); Davies and Ferris (2022).

Human judgement forecasting is not always feasible, especially at scale, due to the time and effort required of human forecasters. However, Human judgement forecasting has various attractive qualities compared to mathematical models. In particular, in the early stages of an outbreak, humans may be able to provide rapid forecasts based on very sparse observational data and can make use of contextual information that is difficult to incorporate into mathematical modelling. They may also help provide guidance on questions that computational models cannot answer, such as whether a state or an organisation like the World Health Organisation will provide assistance to help with an outbreak.

Before the beginning of this PhD only one paper, namely Farrow et al. (2017) for influenza and chikungunya, had directly compared the performance of human judgement forecasts of infectious diseases against predictions made by computational models. Since then, three more papers have examined the performance of human judgement forecasts of COVID-19 in direct comparison to computational models, two of them form part of this PhD (Bosse et al., 2022a; McAndrew et al., 2022b; Bosse et al., 2023b). Results overall suggest that a crowd of human forecasters can achieve performance comparable to that of an ensemble of computational models.

[7] https://gjopen.com

3 Evaluating forecasts using `scoringutils` in `R`

The following Chapter presents `scoringutils`, an R package for evaluating forecasts. It explains the package functionality in detail and gives an overview of how practitioners can use it to evaluate and compare the performance of their forecasts. The package forms the basis for all forecast evaluations conducted in this thesis. `scoringutils` was developed to help address an acute need to understand the quality of the forecasts that were produced to inform the COVID response of public health institutions in the UK and abroad in 2020. It was continuously developed and refined to provide the tools needed to evaluate forecasts in an epidemiological context. In addition to the work presented in this PhD thesis, `scoringutils` also supports and facilitates the evaluations conducted by the US and European Forecast Hubs (Cramer et al., 2022; Sherratt et al., 2022), both of which make use of the package.

The scoring rules implemented in `scoringutils` are mostly not specific for forecasts of infectious diseases. Later, Chapter 5 will go into more detail about how forecasts can be scored in a way that takes the particular characteristics of infectious disease forecasts better into account.

Evaluating Forecasts with scoringutils in R

Nikos I. Bosse
London School of Hygiene & Tropical Medicine (LSHTM)

Hugo Gruson
LSHTM

Anne Cori
Imperial College London

Edwin van Leeuwen
UK Health Security Agency, LSHTM

Sebastian Funk
LSHTM

Sam Abbott
LSHTM

Abstract

Evaluating forecasts is essential to understand and improve forecasting and make forecasts useful to decision makers. A variety of R packages provide a broad variety of scoring rules, visualisations and diagnostic tools. One particular challenge, which **scoringutils** aims to address, is handling the complexity of evaluating and comparing forecasts from several forecasters across multiple dimensions such as time, space, and different types of targets. **scoringutils** extends the existing landscape by offering a convenient and flexible **data.table**-based framework for evaluating and comparing probabilistic forecasts (forecasts represented by a full predictive distribution). Notably, **scoringutils** is the first package to offer extensive support for probabilistic forecasts in the form of predictive quantiles, a format that is currently used by several infectious disease Forecast Hubs. The package is easily extendable, meaning that users can supply their own scoring rules or extend existing classes to handle new types of forecasts. **scoringutils** provides broad functionality to check the data and diagnose issues, to visualise forecasts and missing data, to transform data before scoring, to handle missing forecasts, to aggregate scores, and to visualise the results of the evaluation. The paper presents the package and its core functionality and illustrates common workflows using example data of forecasts for COVID-19 cases and deaths submitted to the European COVID-19 Forecast Hub.

Keywords: forecasting, forecast evaluation, proper scoring rules, scoring, R.

1. Introduction

Good forecasts are of great interest to decision makers in various fields like finance (Timmermann 2018; Elliott and Timmermann 2016), weather predictions (Gneiting and Raftery 2005; Kukkonen *et al.* 2012) or infectious disease modeling (Reich *et al.* 2019; Funk *et al.* 2020; Cramer *et al.* 2021; Bracher *et al.* 2022; Sherratt *et al.* 2022). For decades, researchers, especially in the field of weather forecasting, have therefore developed and refined an arsenal of techniques to evaluate predictions (see for example Good (1952), Epstein (1969); Murphy (1971); Matheson and Winkler (1976), Gneiting, Balabdaoui, and Raftery (2007), Funk, Camacho, Kucharski, Lowe, Eggo, and Edmunds (2019), Gneiting and Raftery (2007), Bracher,

Ray, Gneiting, and Reich (2021)).

Various R (R Core Team 2021) packages cover a wide variety of scoring rules, plots and metrics that are useful in assessing the quality of a forecast. Existing packages offer functionality that is well suited to evaluate a variety of predictive tasks, but also come with important limitations.

Some packages such as **tscount** (Liboschik, Fokianos, and Fried 2017), **topmodels** (Zeileis and Lang 2022), **GLMMadaptive** (Rizopoulos 2023), **cvGEE** (Rizopoulos 2019) or **fabletools** (O'Hara-Wild, Hyndman, and Wang 2023) expect that forecasts were generated in a certain way and require users to supply an object of a specific class to compute scores. These packages provide excellent tools for users operating within the specific package framework but are by their nature not generally applicable to many use cases practitioners might encounter.

Packages such as **scoringRules** (Jordan, Krüger, and Lerch 2019), **Metrics** (Hamner and Frasco 2018), **MLmetrics** (Yan 2016), **verification** (Laboratory 2015), **SpecsVerification** (Siegert 2020), **surveillance** (Meyer, Held, and Höhle 2017), **predtools** (Sadatsafavi, Safari, and Lee 2023), or **probably** (Kuhn, Vaughan, and Ruiz 2023b) provide an extensive collection of tools, scoring rules and visualisations for various use cases. However, most scoring functions operate on vectors and matrices. This is desirable in many applications but can make it difficult to simultaneously evaluate multiple forecasts across several dimensions, such as time, space, and different types of targets.

scoring (Merkle and Steyvers 2013) operates on a data.frame and uses a formula interface, making this task easier. However, **scoring** only exports a few scoring rules and does not allow users to supply their own. **yardstick** (Kuhn, Vaughan, and Hvitfeldt 2023a), which builds on the **tidymodels** (Kuhn and Wickham 2020) framework, is the most general and flexible other forecast evaluation package. It allows users to apply arbitrary scoring rules to a data.frame of forecasts, independently of how they were created. However, **yardstick** is primarily focused on point forecasts and classification tasks. It currently lacks general support for probabilistic forecasts (forecasts in the form of a full predictive distribution, represented e.g. by a set of quantiles or samples from the forecast distribution). Probabilistic forecasts are desirable, as they allow decision makers to take into account the uncertainty of a forecast (Gneiting *et al.* 2007), and are widely used, e.g. in Meteorology or Epidemiology.

scoringutils aims to fill the existing gap in the ecosystem by providing a flexible general-purpose tool for the evaluation of probabilistic forecasts. It offers a coherent `data.table`-based framework and workflow that allows users to evaluate and compare forecasts across multiple dimensions using a wide variety of default and user-provided scoring rules. Notably, **scoringutils** is the first package to offer extensive support for probabilistic forecasts in the form of predictive quantiles, a format that is currently used by several infectious disease Forecast Hubs (Reich *et al.* 2019; Cramer *et al.* 2020; Sherratt *et al.* 2022; Bracher *et al.* 2022). The package provides broad functionality to check the data and diagnose issues, to visualise forecasts and missing data, to transform data before scoring (see Bosse, Abbott, Cori, van Leeuwen, Bracher, and Funk 2023), to apply various metrics and scoring rules to data, to handle missing forecasts, to aggregate scores and to visualise the results of the evaluation. **scoringutils** makes extensive use of **data.table** (Dowle and Srinivasan 2023) to ensure fast and memory-efficient computations. The core functionality is designed around S3 classes, allowing users to expand on the generics and methods implemented in the package. **scoringutils** provides extensive documentation and case studies, as well as sensible defaults

for scoring forecasts.

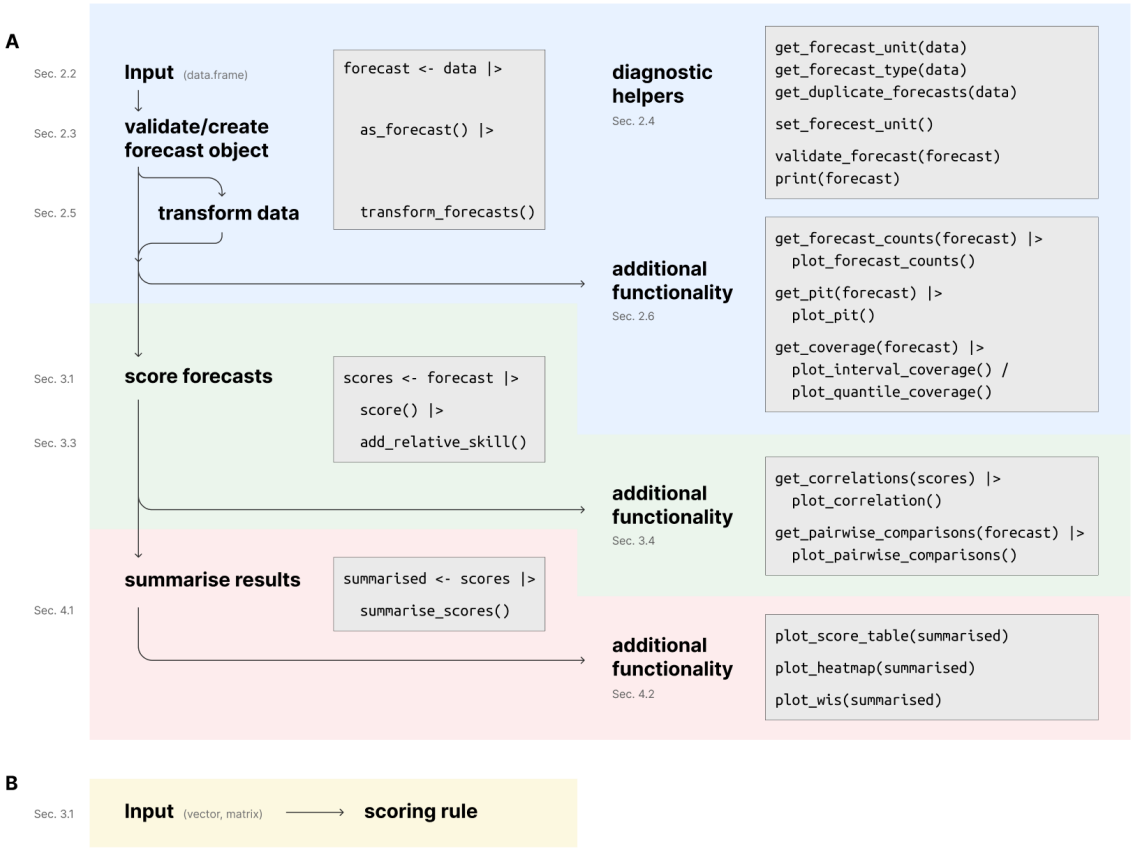

Figure 1: Illustration of the suggested workflow for evaluating forecasts with **scoringutils**. A: Workflow for working with forecasts in a `data.table`-based format. The left side shows the core workflow of the package: 1) validating and processing inputs, 2) scoring forecasts and 3) summarising scores. The right side shows additional functionality that is available at the different stages of the evaluation process. The part in blue is covered by Section 2 and includes all functions related to processing and validating inputs as well as obtaining additional information about the forecasts. The part in green is covered by Section 3 and includes all functions related to scoring forecasts and obtaining additional information about the scores. The part in red is covered by Section 4 and includes all functions related to summarising scores and additional visualisations based on summarised scores. B: An alternative workflow, allowing users to call scoring rules directly with vectors/matrices as inputs.

Paper outline and package workflow

The structure of this paper follows the suggested package workflow which consists of 1) validating and processing inputs, 2) scoring forecasts and 3) summarising scores. This workflow is illustrated in Figure 1, which displays the core workflow (left side) as well as additional functionality that is available at different stages of the evaluation process (right side).

Section 2 is centred around validating inputs, `forecast` objects, and the associated function-

ality. It explains the expected input formats and how to validate inputs and diagnose issues. It provides an overview of the types of forecasts supported by **scoringutils** and the different S3 classes used to represent these forecast types. It also provides information on a variety of functions that can be used to visualise forecasts, transform inputs or obtain additional information and visualisations.

Section 3 is centred around scoring forecasts and the additional functionality that is available to manipulate and analyse scores further. It explains how to score forecasts, either in a `data.table`-format or in a format based on matrices and vectors. It also provides information on additional information that can be computed from scores, such as correlations between scores or relative skill scores based on pairwise comparisons. These can be useful to mitigate the effects of missing forecasts.

Section 4 is centred around summarised scores. It explains how to summarise scores and gives information on additional visualisations that can be created based on summarised scores.

Section 5 discusses the merits and limitations of the package in its current version as explores avenues for future work.

All functionality will be illustrated using the example data shipped with the package, which is based on a subset of case and death forecasts submitted every week between May and September 2021 to the European COVID-19 Forecast Hub (Sherratt *et al.* 2022). Following the convention of the different COVID-19 Forecast Hubs, we will restrict examples to two-week-ahead forecasts.

The code for this package and paper can be found on `https:github.com/epiforecasts/scoringutils`. The full package documentation as well as an overview of all existing functions can also be seen on `https://epiforecasts.io/scoringutils`.

2. Inputs, forecast types and input validation

2.1. Input formats and types of forecasts

Forecasts differ in the exact prediction task and in how the forecaster chooses to represent their prediction. To distinguish different kinds of forecasts, we use the term "forecast type" (which is more a convenient classification than a formal definition). Currently, `scoringutils` distinguishes four different forecast types: "binary", "point", "quantile" and "sample" forecasts.

- "Binary" denotes a probability forecast for a binary (yes/no) outcome variable. This is sometimes also called "soft binary classification".
- "Point" denotes a forecast for a continuous or discrete outcome variable that is represented by a single number.
- "Quantile" or "quantile-based" is used to denote a probabilistic forecast for a continuous or discrete outcome variable, with the forecast distribution represented by a set of predictive quantiles. While a single quantile would already satisfy the requirements for a quantile-based forecast, most scoring rules expect a set of quantiles which are symmetric around the median (thus forming the lower and upper bounds of central "prediction intervals") and will return `NA` if this is not the case.

- "Sample" or "sample-based" is used to denote a probabilistic forecast for a continuous or discrete outcome variable, with the forecast represented by a finite set of samples drawn from the predictive distribution. A single sample technically suffices, but would lead to very imprecise results.

Forecast type			**column**	**type**

All forecast types			`observed`
			`predicted`
			`model`	

Classification	Binary	Soft classification (prediction is probability)	`observed` `predicted`	factor with 2 levels numeric [0,1]
Point forecast			`observed` `predicted`	numeric numeric

Probabilistic forecast		Sample format	`observed`
			`predicted`
			`sample_id`
				numeric
				numeric
				numeric

		Quantile format	`observed`
			`predicted`
			`quantile_level`
				numeric
				numeric
				numeric [0,1]

Table 1: Formatting requirements for data inputs. Regardless of the forecast type, the `data.frame` (or similar) must have columns called `observed`, `predicted`, and `model`. For binary forecasts, the column `observed` must be of type factor with two levels and the column `predicted` must be a numeric between 0 and 1. For all other forecast types, both `observed` and `predicted` must be of type numeric. Forecasts in a sample-based format require an additional numeric column `sample_id` and forecasts in a quantile-based format require an additional numeric column `quantile_level` with values between 0 and 1.

The starting point for working with **scoringutils** is usually a `data.frame` (or similar) containing both the predictions and the observed values. In a next step (see Section 2.2) this data will be validated and transformed into a "forecast object". The input data needs to have a column `observed` for the observed values, a column `predicted` for the predicted values, and a column `model` denoting the name of the model/forecaster that generated the forecast. Additional requirements depend on the forecast type. Table 1 shows the expected input format for each forecast type.

The package contains example data for each forecast type, which can serve as an orientation for the correct formats. The example data sets are exported as `example_quantile`, `example_continuous`, `example_integer`, `example_point` and `example_binary`. For illustrative purposes, the example data also contains some rows with only observations and no corresponding predictions. Input formats for the scoring rules that can be called directly follow the same convention, with inputs expected to be vectors or matrices.

The unit of a single forecast

Apart from the columns `observed`, `predicted`, `model`, and the extra columns required for each forecast type, it is usually necessary that the input data contains additional columns.

This is because a single probabilistic forecast (apart from binary predictions) is composed of multiple values. A quantile-based forecast, for example, is composed of several quantiles, and a sample-based forecast of multiple samples. However, every row only holds a single sample/quantile. Several rows in the input data therefore jointly form a single forecast. Additional columns in the input provide the information necessary to group rows that belong to the same forecast. The combination of values in those columns forms the unit of a single forecast (or "forecast unit") and should uniquely identify a single forecast. For example, consider forecasts made by different models in various locations at different time points and for different targets. A single forecast could then be uniquely described by the values in the columns `model`, `location`, `date`, and `target`, and the forecast unit would be `forecast_unit = c("model", "location", "date", "target")`.

Rows are automatically grouped based on the values in all other columns present in the data (excluding required columns like `sample_id` or `quantile_level` and values computed by **scoringutils**). As the forecast unit is determined based on all existing columns, no column must be present that is unrelated to the forecast unit. As a very simplistic example, consider an additional row, `"even"`, that is one if the row number is even and zero otherwise. The existence of this column would change results, as **scoringutils** assumes it was relevant to grouping the forecasts.

2.2. Forecast objects and input validation

The raw input data needs to be processed and validated using the function `as_forecast()`:

```
R> library(scoringutils)
R> forecast_quantile <- example_quantile[horizon == 2] |>
+   as_forecast()
```

The function `as_forecast()` recognises the type of the forecast based on the available columns, transforms the input into a "forecast" object and validates it (see Figure A.11 for details). A forecast object is a `data.table` that has passed some input validations. It behaves like a `data.table`, but has dedicated methods e.g. for input validation, scoring and printing. The classes corresponding to the forecast types are `forecast_point`, `forecast_binary`, `forecast_quantile` and `forecast_sample`.

`as_forecast()` can automatically determine the forecast type and forecast unit based on the input data. However, it can also take additional arguments that help facilitate the process of creating a forecast object:

```
R> forecast_quantile <- example_quantile[horizon == 2] |>
+   as_forecast(
+     forecast_unit = c(
+       "model", "location", "target_end_date",
+       "forecast_date", "horizon", "location"
+     ),
+     forecast_type = "quantile",
+     observed = "observed",
+     predicted = "predicted",
```

```
+     model = "model",
+     quantile_level = "quantile_level",
+   )
```

The argument `forecast_unit` allows the user to manually set the unit of a single forecast. This is done by dropping all columns that are not either specified in the `forecast_unit` or are "protected" columns (such as `observed`, `predicted`, `model`, `quantile_level`, or `sample_id`). The argument `forecast_type` allows users to manually specify the forecast type they expect. If the forecast type inferred from the input does not match the specified forecast type, an error is thrown. The other arguments can be used to specify the column names of the input data that correspond to the required columns. `as_forecast()` will rename the specified columns to the corresponding required columns.

2.3. Diagnostic helper functions

Various helper functions are available to diagnose and fix issues with the input data. The most important one is `print()`. Once a forecast object has successfully been created, diagnostic information will automatically be added to the output when printing a forecast object. This information includes the forecast type, the forecast unit, and additional information in case the object fails validations.

```
R> print(forecast_quantile, 2)

Forecast type:
[1] "quantile"

Forecast unit:
[1] "location"        "target_end_date" "target_type"
[4] "location_name"   "forecast_date"   "model"
[7] "horizon"

Key: <location, target_end_date, target_type>
       location target_end_date target_type observed location_name
         <char>          <Date>      <char>    <num>        <char>
    1:       DE      2021-01-02       Cases   127300       Germany
    2:       DE      2021-01-02      Deaths     4534       Germany
   ---
20544:       IT      2021-07-24      Deaths       78         Italy
20545:       IT      2021-07-24      Deaths       78         Italy
       forecast_date quantile_level predicted                model
              <Date>          <num>     <int>               <char>
    1:          <NA>             NA        NA                 <NA>
    2:          <NA>             NA        NA                 <NA>
   ---
20544:    2021-07-12          0.975       611 epiforecasts-EpiNow2
20545:    2021-07-12          0.990       719 epiforecasts-EpiNow2
       horizon
```

```
        <num>
    1:      NA
    2:      NA
   ---
20544:       2
20545:       2
```

Internally, the print method calls the functions `get_forecast_type()`, `get_forecast_unit()` and `validate_forecast()`. `get_forecast_type()` and `get_forecast_unit()` work on either an unvalidated `data.frame` (or similar) or on an already validated forecast object. They return the forecast type and the forecast unit, respectively, as inferred from the input data. `validate_forecast()` re-validates an existing forecast object an can be used programmatically without printing an object (users could in principle also call `as_forecast()` again).

One common issue that causes `as_forecast()` to fail are "duplicates" in the data. **scoringutils** strictly requires that there be only one forecast per forecast unit and only one predicted value per quantile level or sample id within a single forecast. Duplicates usually occur if the forecast unit is misspecified. For example, if we removed the column `target_type` from the example data, we would now have two forecasts (one for cases and one for deaths of COVID-19) that appear to have the same forecast unit (since the information that distinguished between case and death forecasts is no longer there). The function `get_duplicate_forecasts()` returns duplicate rows for the user to inspect. To remedy the issue, the user needs to add additional columns that uniquely identify a single forecast.

```
R> rbind(example_quantile, example_quantile[1001:1002]) |>
+   get_duplicate_forecasts()
```

```
   location target_end_date target_type observed location_name
     <char>          <Date>      <char>    <num>        <char>
1:       DE      2021-05-22      Deaths     1285       Germany
2:       DE      2021-05-22      Deaths     1285       Germany
3:       DE      2021-05-22      Deaths     1285       Germany
4:       DE      2021-05-22      Deaths     1285       Germany
   forecast_date quantile_level predicted                model
          <Date>          <num>     <int>               <char>
1:    2021-05-17          0.975      1642 epiforecasts-EpiNow2
2:    2021-05-17          0.990      1951 epiforecasts-EpiNow2
3:    2021-05-17          0.975      1642 epiforecasts-EpiNow2
4:    2021-05-17          0.990      1951 epiforecasts-EpiNow2
   horizon
     <num>
1:       1
2:       1
3:       1
4:       1
```

2.4. Transforming forecasts

As suggested in Bosse *et al.* (2023), users may want to transform forecasts before scoring them. Two commonly used scoring rules are the continuous ranked probability score (CRPS) and the weighted interval score (WIS). Both measure the absolute distance between the forecast and the observation. This may not be desirable, for example in the context of epidemiological forecasts, where infectious disease processes are usually modelled to occur on a multiplicative scale. Taking the logarithm of the forecasts and observations before scoring them makes it possible to evaluate forecasters based on how well they predicted the exponential growth rate.

The function `transform_forecasts()` takes a validated forecast object as input and allows users to apply arbitrary transformations to forecasts and observations. Users can specify a function via the argument `fun` (as well as supply additional function parameters). The default function is `log_shift()`, which is simply a wrapper around `log()` with an additional argument that allows adding an offset (i.e. `log(x + offset)`) to deal with zeroes in the data. Users can specify to either append the transformed forecasts to the existing data by setting `append = TRUE` (the default behaviour, resulting in an additional column `scale`) or to replace the existing forecasts in place.

The example data contains negative values which need to be handled before applying the logarithm. Presumably, negative values for count data should be dropped altogether, but for illustrative purposes, we will call `transform_forecasts()` twice to replace them with zeroes first before appending transformed counts.

```
R> forecast_quantile |>
+   transform_forecasts(fun = \(x) {pmax(x, 0)}, append = FALSE) |>
+   transform_forecasts(fun = log_shift, offset = 1) |>
+   print(2)
```

```
       location target_end_date target_type     observed
         <char>          <Date>      <char>        <num>
    1:       DE      2021-01-02       Cases 1.273000e+05
    2:       DE      2021-01-02      Deaths 4.534000e+03
   ---
41089:       IT      2021-07-24      Deaths 4.369448e+00
41090:       IT      2021-07-24      Deaths 4.369448e+00
       location_name forecast_date quantile_level predicted
              <char>        <Date>          <num>     <num>
    1:       Germany          <NA>             NA        NA
    2:       Germany          <NA>             NA        NA
   ---
41089:         Italy    2021-07-12          0.975  6.416732
41090:         Italy    2021-07-12          0.990  6.579251
                     model horizon   scale
                    <char>   <num>  <char>
    1:                <NA>      NA natural
    2:                <NA>      NA natural
   ---
41089: epiforecasts-EpiNow2       2     log
41090: epiforecasts-EpiNow2       2     log
```

2.5. Additional functionality related to forecast objects

scoringutils offers a variety of different functions that allow users to obtain and visualise additional information about their forecast. The package also has an extensive Vignette with examples for further visualisations that are not implemented as functions.

Displaying the number of forecasts available

Users can get an overview of how many forecasts there are using `get_forecast_counts()`. The function takes a validated forecast object as input and returns a data.table of forecast counts, which helps obtain an overview of missing forecasts. This can impact the evaluation, if missingness correlates with performance. Users can specify the level of summary through the `by` argument. For example, to see how many forecasts there are per `model`, `target_type` and `forecast_date`, we can run

```
R> forecast_counts <- forecast_quantile |>
+   get_forecast_counts(
+     by = c("model", "target_type", "forecast_date")
+   )
```

We can visualise the results by calling `plot_forecast_counts()` on the output (Figure 2).

```
R> library(ggplot2)
R> forecast_counts |>
+   plot_forecast_counts(x = "forecast_date") +
+   facet_wrap(~ target_type) +
+   labs(y = "Model", x = "Forecast date")
```

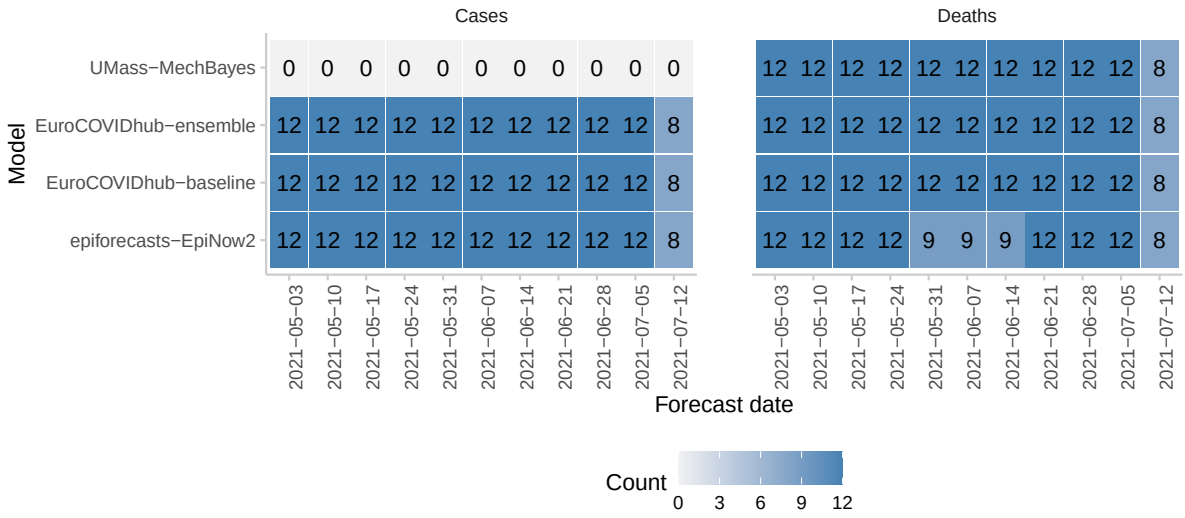

Figure 2: Visualistion of forecast counts for the example data. Numbers (and colour shade) indicate the number of forecasts available for a given model, target type and forecast date.

Probabilistic calibration and PIT histograms

One important quality of good forecasts is calibration. The term describes a statistical consistency between the forecasts and the observations, i.e. an absence of systematic deviations between the two. It is possible to distinguish several forms of calibration which are discussed in detail by Gneiting *et al.* (2007). The form of calibration most commonly focused on is called probabilistic calibration. Probabilistic calibration means that the forecast distributions are consistent with the true data-generating distributions in the sense that on average, τ% of true observations will be below the corresponding τ-%-quantiles of the cumulative forecast distributions.

A common way to visualise probabilistic calibration is the probability integral transform (PIT) histogram (Dawid 1984). Observed values, y, are transformed using the CDF of the predictive distribution, F, to create a new variable u with $u = F(y)$. u is therefore simply the CDF of the predictive distribution evaluated at the observed value. If forecasts are probabilistically calibrated, then the transformed values will be uniformly distributed (for a proof see for example Angus (1994)). When plotting a histogram of PIT values (see Figure 3), a systematic bias usually leads to a triangular shape, a U-shaped histogram corresponds to forecasts that are underdispersed (too sharp) and a hump shape appears when forecasts are overdispersed (too wide). There exist different variations of the PIT to deal with discrete instead of continuous data (see e.g. Czado, Gneiting, and Held (2009) and Funk *et al.* (2019)). The PIT version implemented in `scoringutils` for discrete variables follows Funk *et al.* (2019).

Users can obtain PIT histograms based on validated forecast objects using the function `get_pit()` and can visualise results using `plot_pit()`. Once again, the argument `by` controls the summary level. The output of the following is shown in Figure 3:

```
R> example_continuous |>
+   get_pit(by = c("model", "target_type")) |>
+   plot_pit() +
+   facet_grid(target_type ~ model)
```

It is, in theory, possible to conduct a formal test for probabilistic calibration, for example by employing an Anderson-Darling test on the uniformity of PIT values. In practice, this can be difficult as forecasts, and therefore PIT values as well, are often correlated. Personal experience suggests that the Anderson-Darling test is often too quick to reject the null hypothesis of uniformity. It is also important to note that uniformity of the PIT histogram does not guarantee that forecasts are indeed calibrated. Gneiting *et al.* (2007); Hamill (2001) provide examples with different forecasters who are mis-calibrated, but have uniform PIT histograms.

Probabilistic calibration and coverage plots

For forecasts in a quantile-based format, there exists a second way to assess probabilistic calibration: we can easily compare the proportion of observations that fall below the τ-quantiles of all forecasts ("empirical quantile coverage") to the nominal quantile coverage τ. Similarly, we can compare the empirical coverage of the central prediction intervals formed by the predictive quantiles to the nominal interval coverage. For example, the central 50% prediction intervals of all forecasts should contain around 50% of the observed values, the 90% central intervals should contain around 90% of observations etc. In addition, we can define coverage deviation as the difference between nominal and empirical coverage.

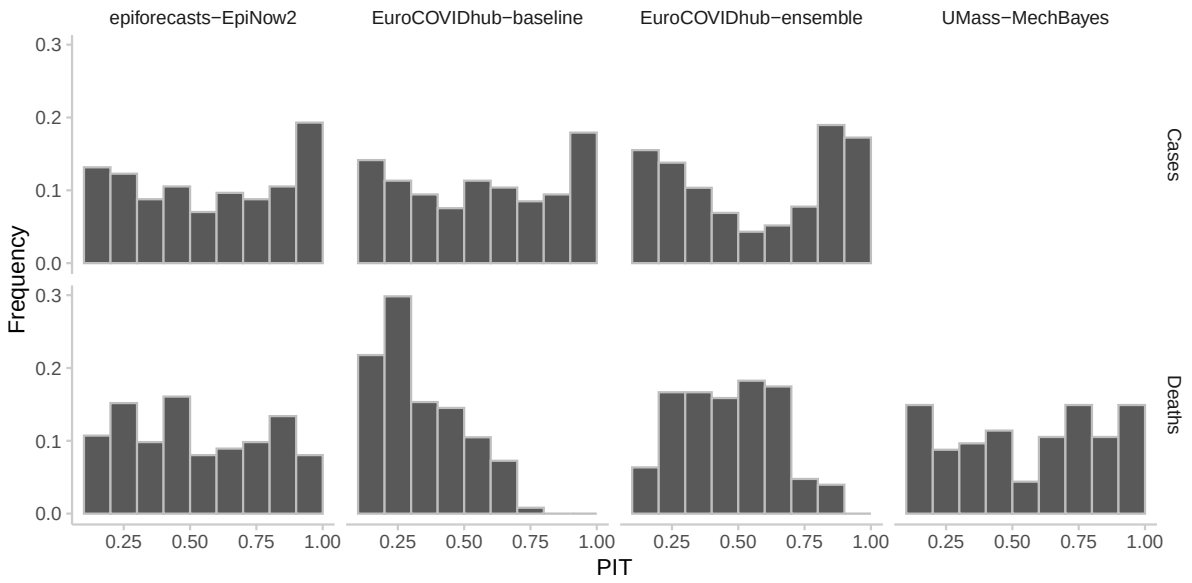

Figure 3: PIT histograms of all models stratified by forecast target. Histograms should ideally be uniform. A u-shape usually indicates overconfidence (forecasts are too narrow), a hump-shaped form indicates underconfidence (forecasts are too uncertain) and a triangle-shape indicates bias.

Interval and quantile coverage can easily be computed by calling `get_coverage()` on a validated forecast object (in a quantile-based format). The function computes interval coverage, quantile coverage, interval coverage deviation and quantile coverage deviation and returns a `data.table` with corresponding columns. Coverage values will be summarised according to the level specified in the `by` argument and one value per quantile level/interval range is returned.

```
R> forecast_quantile |>
+   get_coverage(by = "model") |>
+   print(2)
```

Results can then be visualised using the functions `plot_interval_coverage()` (see Figure 4A) and `plot_quantile_coverage()` (see 4B). Both show nominal against empirical coverage. Ideally, forecasters should lie on the diagonal line. If the line moves into the green-shaded area, the forecaster is too conservative, i.e. the predictive distributions are too wide/overdispersed on average. The white area implies overconfidence/predictive distributions that are too narrow on average (see Figure B.12) for more details).

```
R> coverage <- get_coverage(forecast_quantile, by = c("model", "target_type"))
R>
R> plot_interval_coverage(coverage) +
+   facet_wrap(~ target_type)
R>
R> plot_quantile_coverage(coverage) +
+   facet_wrap(~ target_type)
```

Note that users can also compute individual coverage values as scores using `score()`. This represents a separate workflow that allows users to obtain coverage values as a summary measure to be computed alongside other scores, rather than providing a way to visually assess calibration.

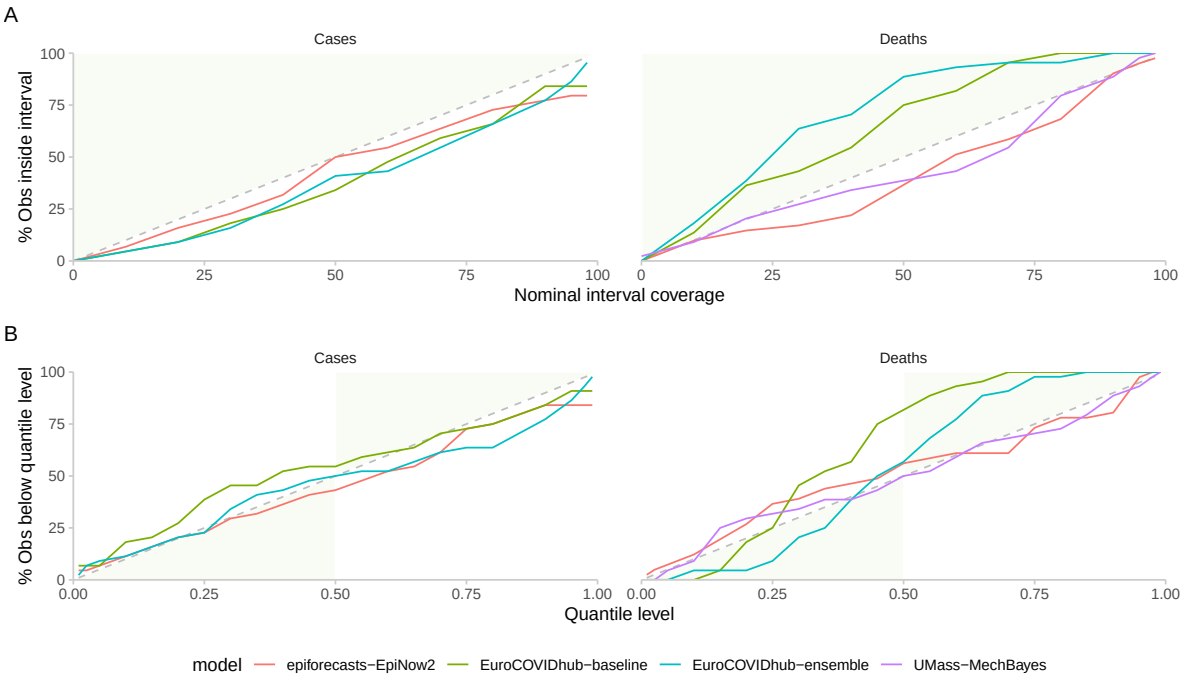

Figure 4: Interval coverage (A) and quantile coverage (B) plots. Areas shaded in green indicate that the forecasts are too wide (i.e., underconfident), while areas in white indicate that the model is overconfident and generates too narrow prediction intervals.

3. Scoring forecasts

Metrics and scoring rules can be applied to data in two different ways: They can be conveniently applied to a data set of observed and predicted values using `score()`, or they be called directly on a set of vectors and matrices. This section will mostly focus on `score()`.

3.1. score() and working with scoring rules

The function `score()` is the workhorse of the package and applies a set of metrics and scoring rules to predicted and observed values. It is a generic function that dispatches to different methods depending on the class of the input. The input of `score()` is a validated forecast object and its output is an object of class `scores`, which is a essentially `data.table` with an additional attribute `metrics` (containing the names of the metrics used for scoring).

```
R> example_point[horizon == 2] |>
+   as_forecast() |>
+   score() |>
+   print(2)
```

```
Key: <location, target_end_date, target_type>
     location target_end_date target_type observed location_name
       <char>          <Date>      <char>    <num>        <char>
  1:       DE      2021-05-15       Cases    64985       Germany
  2:       DE      2021-05-15       Cases    64985       Germany
 ---
304:       IT      2021-07-24      Deaths       78         Italy
305:       IT      2021-07-24      Deaths       78         Italy
     forecast_date predicted                 model horizon ae_point
            <Date>     <int>                <char>   <num>    <num>
  1:    2021-05-03    110716 EuroCOVIDhub-ensemble       2    45731
  2:    2021-05-03    132607 EuroCOVIDhub-baseline       2    67622
 ---
304:    2021-07-12       124       UMass-MechBayes       2       46
305:    2021-07-12       186  epiforecasts-EpiNow2       2      108
       se_point       ape
          <num>     <num>
  1: 2091324361 0.7037162
  2: 4572734884 1.0405786
 ---
304:       2116 0.5897436
305:      11664 1.3846154
```

All `score()` methods take an argument `metrics` with a named list of functions to apply to the data. These can be metrics exported by **scoringutils** or any other custom scoring function. All metrics scoring rules passed to `score()` need to adhere to the same input format (see Figure 5), corresponding to the type of forecast to be scored. Scoring functions must accept a vector of observed values as their first argument, a matrix/vector of predicted values as their second argument and, for quantile-based forecasts, a vector of quantile levels as their third argument). However, functions may have arbitrary argument names. Within `score()`, inputs like the observed and predicted values, quantile levels etc. are passed to the individual scoring rules by position, rather than by name. The default scoring rules for point forecasts, for example, comprise functions from the **Metrics** package, which use the names `actual` and `predicted` for their arguments instead of `observed` and `predicted`. Additional arguments can be passed down to the scoring functions via the `...` arguments in `score()`.

Composing a custom list of metrics and scoring rules

For every forecast type, there exists a default list of scoring rules that are applied to the data when calling `score()`. The default lists can be accessed by calling the functions `metrics_point()`, `metrics_binary()`, `metrics_sample()` and `metrics_quantile()`. These functions take additional arguments `exclude` and `select` which can be used to customise which scoring rules are included. Alternatively, users can call the function `select_metrics()` on a list of scoring rules, which achieves the same purposes and allows users to compose custom lists of metrics and scoring rules.

```
R> custom_metrics <- metrics_quantile() |>
+   select_metrics(select = c("wis", "overprediction"))
```

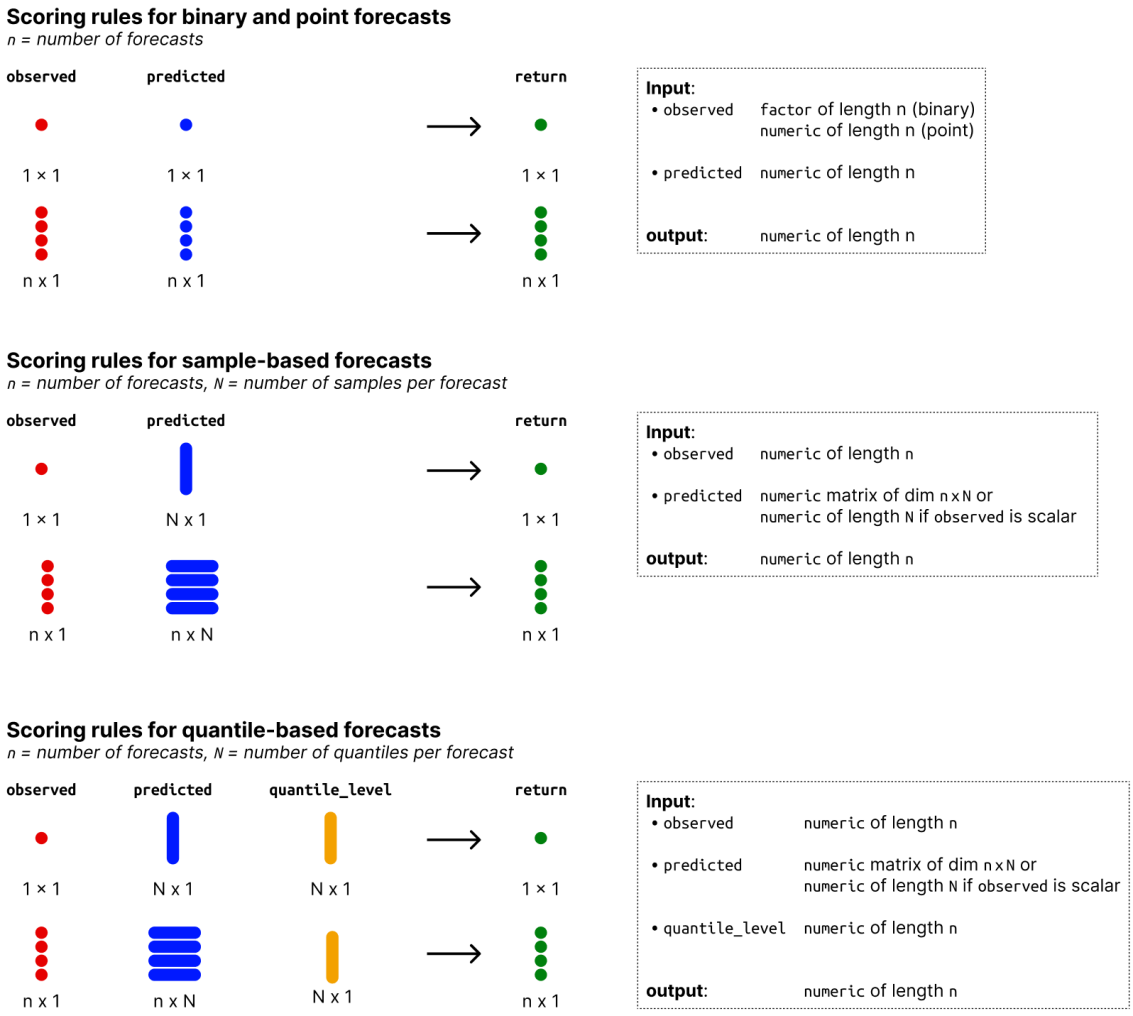

Figure 5: Overview of the inputs and outputs of the metrics and scoring rules exported by **scoringutils**. Dots indicate scalar values, while bars indicate vectors (comprised of values that belong together). Several bars (vectors) can be grouped into a matrix with rows representing the individual forecasts. All scoring functions used within `score()` must accept the same input formats as the functions here. However, functions used within `score()` do not necessarily have to have the same argument names (see Section 3). Input formats directly correspond to the required columns for the different forecast types (see Table 1). The only exception is the forecast type 'sample': Inputs require a column `sample_id` in `score()`, but no corresponding argument is necessary when calling scoring rules directly on vectors or matrices.

```
R>
R> score(metrics = custom_metrics)
```

Details on metrics exported by scoringutils

All metrics are named according to the following schema: `{metric name}_{forecast type}`.

If only a single forecast type is possible, then `_{forecast type}` is omitted. The return value is a vector with scores (only in the case of `wis()`, which is composed of three components (see C), is there an optional argument that causes the function to return a list of vectors for the individual WIS components). The first argument of all metrics exported by **scoringutils** is always `observed`, and the second one is `predicted`. Scoring rules for quantile-based forecasts have an additional argument, `quantile_level`, to denote the quantile levels of the predictive quantiles.

Metrics exported by **scoringutils** differ in the relationship between input and output. Some scoring rules have a one-to-one relationship between predicted values and scores, returning one value per value in `predicted`. This is the case for all metrics for binary and point forecasts. Other scoring rules have a many-to-one relationship, returning one value per multiple values in `predicted`. This is the case for all scoring rules for sample- and quantile-based forecasts. For sample- and quantile-based forecasts, `predicted` is therefore a matrix, with values in each row jointly forming a single forecast.

Input formats and return values are shown in more detail in Figure 5. The package vignettes provide extensive documentation for the metrics exported by **scoringutils** and offer guidance on which scoring rule to use and how to interpret the scores.

3.2. Adding relative skill scores based on pairwise comparisons

Raw scores for different forecasting models are usually not directly comparable when there are missing forecasts in the data set, as missingness is often correlated with predictive performance. One way to mitigate this are relative skill scores based on pairwise comparisons (Cramer *et al.* 2021).

Models enter a 'pairwise tournament', where all possible pairs of models are compared based on the overlapping set of available forecasts common to both models (omitting comparisons where there is no overlapping set of forecasts). For every pair, the ratio of the mean scores of both models is computed. The relative skill score of a model is then the geometric mean of all mean score ratios which involve that model (see Figure 6. This gives us an indicator of performance relative to all other models, with the orientation depending on the score used: if lower values are better for a particular scoring rule, then the same is true for the relative skill score computed based on that score.

Two models can of course only be fairly compared if they have overlapping forecasts. Furthermore, pairwise comparisons between models for a given score are only possible if all values have the same sign, i.e. all score values need to be either positive or negative.

To compute relative skill scores, users can call `add_pairwise_comparison()` on the output of `score()`. This function computes relative skill values with respect to a score specified in the argument `metric` and adds them as an additional column to the input data. Optionally, users can specify a baseline model to also compute relative skill scores scaled with respect to that baseline. Scaled relative skill scores are obtained by simply dividing the relative skill score for every individual model (computed excluding the baseline) by the relative skill score of the baseline model. Pairwise comparisons are computed according to the grouping specified in the argument `by`: internally, the `data.table` with all scores gets split into different `data.table`s according to the values specified in `by` (excluding the column 'model'). Relative scores are then computed for every individual group separately. In the example below we specify `by = c("model", "target_type")`, which means that there is one relative skill score per model,

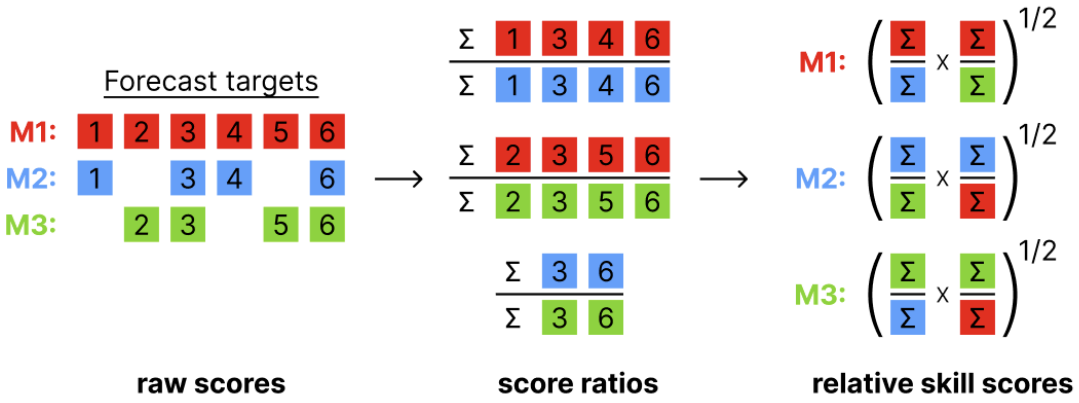

Figure 6: Illustration of the computation of relative skill scores through pairwise comparisons of three different forecast models, M1-M3. Score ratios are computed based on the overlapping set of forecasts common to all pairs of two models. The relative skill score of a model is then the geometric mean of all mean score ratios which involve that model. The orientation of the relative skill score depends on the score used: if lower values are better for a particular scoring rule, then the same is true for the relative skill score computed based on that score.

calculated completely separately for the different forecasting targets.

```
R> forecast_quantile |>
+   score() |>
+   add_relative_skill(by = c("model", "target_type"),
+                      baseline = "EuroCOVIDhub-baseline")
```

Pairwise comparisons should usually be made based on raw, unsummarised scores (meaning that `add_relative_skill()` should be called before `summarise_scores()` (see Section 4)). Summarising scores, for example by computing an average across several dimensions, can change the set of overlapping forecasts between two models and distort relative skill scores.

3.3. Additional functionality related to scores objects

Displaying mean score ratios from pairwise comparisons

scoringutils offers a second alternative workflow to conduct pairwise comparisons between models through the function `get_pairwise_comparisons()`. The purpose of this workflow is to obtain and visualise information on the direct comparisons between every possible pair of models, rather than just computing relative skill scores for every model. The function `get_pairwise_comparisons()` accepts the same inputs as `add_relative_skill()`, and returns a `data.table` with the results of the pairwise tournament. These include the mean score ratios for every pair of models, a p-value for whether scores for one model are significantly different from scores for another model, and the relative and scaled relative skill score for every model (depending on whether a baseline was provided or not).

`get_pairwise_comparisons()` computes p-values using either the Wilcoxon rank sum test (the default, the test is also known as Mann-Whitney-U test) (Mann and Whitney 1947) or a permutation test. P-values are then adjusted using `p.adjust`. In practice, the computation of p-values is complicated by the fact that both tests assume independent observations. In reality, however, forecasts by a model may be correlated across time or space (e.g., if a forecaster has a bad day, they might perform badly across different targets for a given forecast date). P-values may therefore be too liberal in suggesting significant differences where there aren't any. We previously suggested computing relative skill scores based on pairwise comparisons before summarising scores. One exception is the case where one is interested in p-values specifically: One possible way to mitigate issues from correlated forecasts, is to aggregate observations over a category where one suspects correlation (provided there are no missing values within the categories summarised over) to reduce correlation before making pairwise comparisons. A test that is performed on aggregate scores will likely be more conservative.

The mean score ratios resulting from `pairwise_comparison()` can then be visualised using the function `plot_pairwise_comparison()`. An example is shown in Figure 7.

```
R> forecast_quantile |>
+   score() |>
+   get_pairwise_comparisons(by = c("model", "target_type")) |>
+   plot_pairwise_comparisons() +
+   facet_wrap(~ target_type)
```

Correlations between scores

Users can examine correlations between scores using the function `correlations()` and plot the result using `plot_correlations()`. The plot resulting from the following code is shown in Figure 8.

```
R> correlations <- forecast_quantile |>
+   score() |>
+   summarise_scores() |>
+   get_correlations(digits = 2)
R>
R> correlations |>
+   plot_correlations()
```

4. Summarising results

4.1. Summarising scores

Usually, one will not be interested in scores for each individual forecast, but rather in summarised scores. This can be achieved using the function `summarise_scores()`. The function takes a `scores` object (a data.table with an additional attribute `metrics`) as input and applies a summary function to it (by default the mean), returning a `data.table` with summarised scores. Users can set the summary level using the argument `by` and will obtain

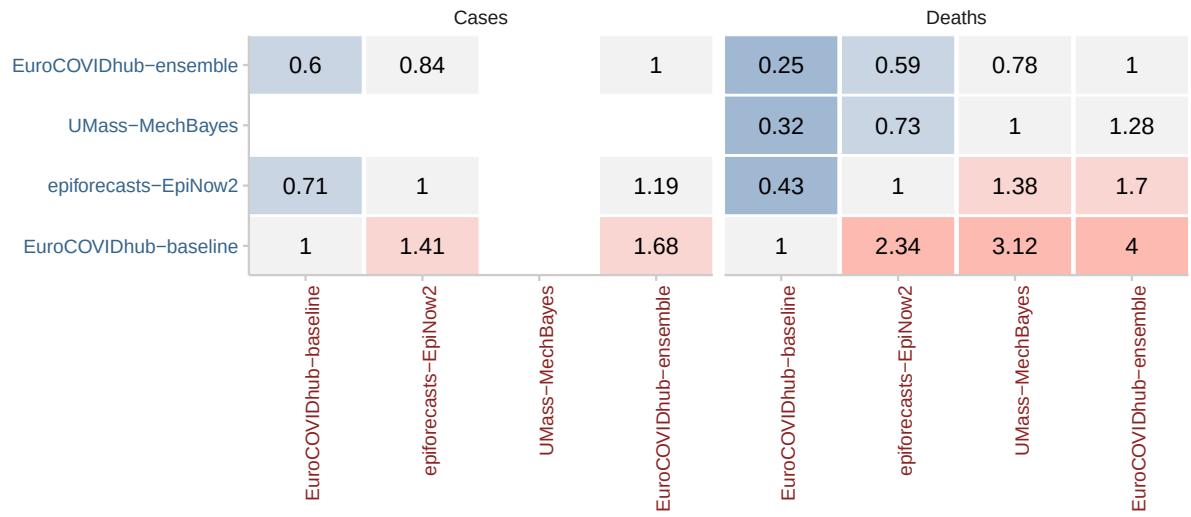

Figure 7: Ratios of mean weighted interval scores based on overlapping forecast sets. When interpreting the plot one should look at the model on the y-axis, and the model on the x-axis is the one it is compared against. If a tile is blue, then the model on the y-axis performed better (assuming that scores are negatively oriented, i.e. that lower scores are better). If it is red, the model on the x-axis performed better in direct comparison. In the example above, the EuroCOVIDhub-ensemble performs best (it only has values smaller than one), while the EuroCOVIDhub-baseline performs worst (and only has values larger than one). For cases, the UMass-MechBayes model is excluded as there are no case forecasts available and therefore the set of overlapping forecasts is empty.

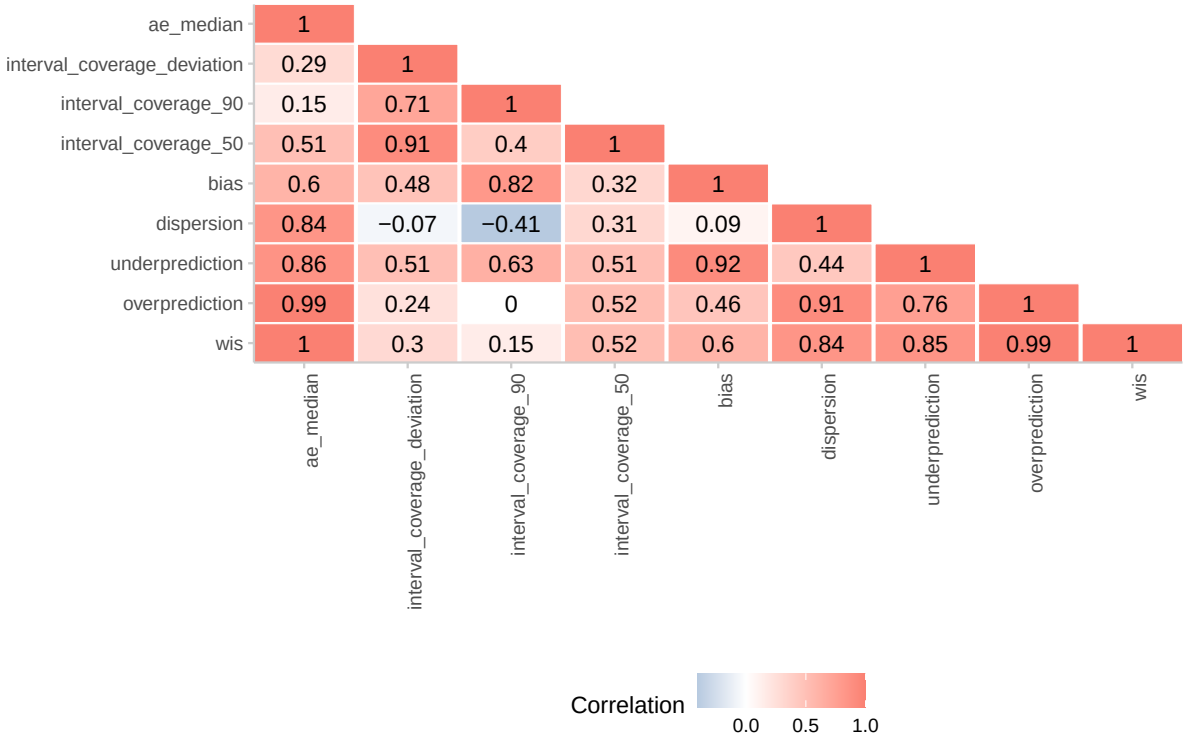

Figure 8: Plot of correlations between different scores. Numbers, as well as the shade of the cells, indicate the correlation between two scores.

a summarised score for each combination of the value in the specified columns (e.g. `by = c("model", "target_type")` will return one summarised score per model and target type). Equivalently, users can specify the columns that should be aggregated over (using the argument `across`). To display scores it is often useful to round the output, for example to two significant digits, which can be achieved with another call to `summarise_scores()`.

```
R> forecast_quantile |>
+   score(metrics = list("wis" = wis)) |>
+   summarise_scores(by = c("model", "target_type")) |>
+   summarise_scores(fun = signif, digits = 2)

                  model   wis
                 <char> <num>
1: EuroCOVIDhub-ensemble 17000
2: EuroCOVIDhub-ensemble    41
3: EuroCOVIDhub-baseline 29000
4: EuroCOVIDhub-baseline   160
5:  epiforecasts-EpiNow2 21000
6:  epiforecasts-EpiNow2    69
7:       UMass-MechBayes    52
```

While `summarise_scores()` accepts arbitrary summary functions, care has to be taken when using something else than `mean()`, as this may create an incentive for dishonest reporting.

Many scoring rules for probabilistic forecasts are 'strictly proper scoring rules' (Gneiting and Raftery 2007), meaning that they are constructed such that they cannot be cheated and always incentivise the forecaster to report her honest belief about the future. Let's assume that a forecaster's true belief about the future corresponds to a predictive distribution F. Then, if F was the true data-generating process, a scoring rule would be proper if it ensures that no other forecast distribution G would yield a better expected score. If the scoring rule ensures that under F no other possible predictive distribution can achieve the same expected score as F, then it is called strictly proper. From the forecaster's perspective, any deviation from her true belief F leads to a worsening of expected scores. When using summary functions other than the mean, however, scores may lose their propriety (the property of incentivising honest reporting) and become cheatable. For example, the median of several individual scores (individually based on a strictly proper scoring rule) is usually not proper. A forecaster judged by the median of several scores may be incentivised to misrepresent their true belief in a way that is not true for the mean score.

The user must exercise additional caution and should usually avoid aggregating scores across categories which differ much in the magnitude of the quantity to forecast, as (depending on the scoring rule used) forecast errors usually increase with the order of magnitude of the forecast target. In the given example, looking at one score per model (i.e., specifying `by = c("model")`) is problematic, as overall aggregate scores would be dominated by case forecasts, while performance on deaths would have little influence. Similarly, aggregating over different forecast horizons is often ill-advised as the mean will be dominated by further ahead forecast horizons. In the previous function calls, we therefore decided to only analyse forecasts with a forecast horizon of two weeks.

4.2. Additional functionality for summarised scores

Heatmaps

To detect systematic patterns it may be useful to visualise a single score across several dimensions. The function `plot_heatmap()` can be used to create a heatmap that achieves this. The following produces a heatmap of bias values across different locations and forecast targets (output shown in Figure 9).

```
R> example_continuous[horizon == 2] |>
+   as_forecast() |>
+   score() |>
+   summarise_scores(by = c("model", "location", "target_type")) |>
+   plot_heatmap(x = "location", metric = "bias") +
+     facet_wrap(~ target_type)
```

Weighted interval score decomposition

For quantile-based forecasts, the weighted interval score (WIS, Bracher *et al.* 2021, see Section C in the Appendix) is a commonly used strictly proper scoring rule for forecasts in a quantile-based format. The score is the sum of three components: overprediction, underprediction and dispersion (width of the forecast). These can be visualised using the function `plot_wis()`, as shown in Figure 10.

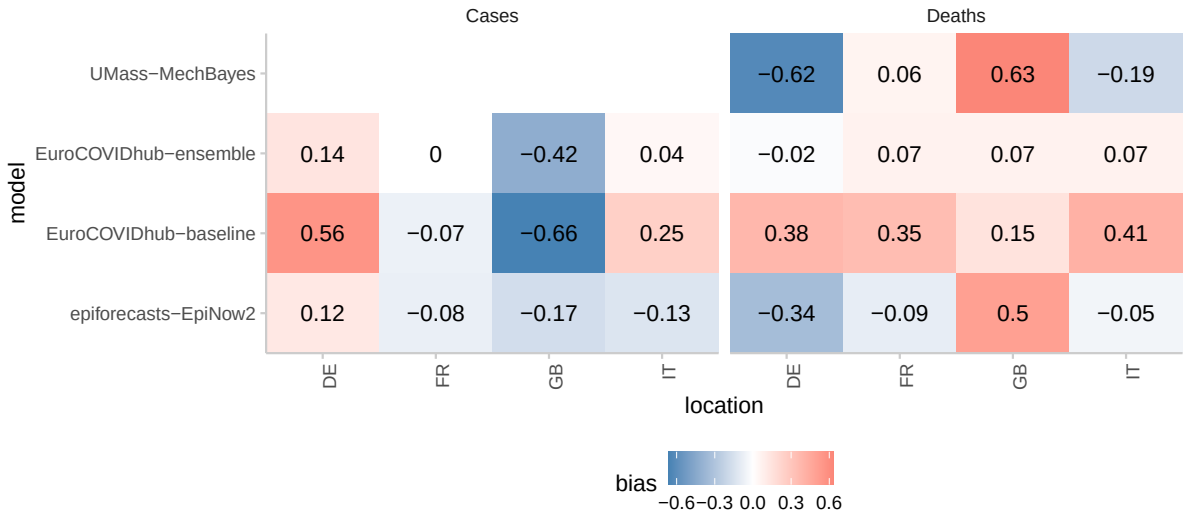

Figure 9: Heatmap of bias values for different models across different locations and forecast targets. Bias values are bound between -1 (underprediction) and 1 (overprediction) and should be 0 ideally. Red tiles indicate an upwards bias (overprediction), while blue tiles indicate a downwards bias (underprediction)

```
R> forecast_quantile |>
+   score() |>
+   summarise_scores(by = c("model", "target_type")) |>
+   plot_wis(relative_contributions = FALSE) +
+   facet_wrap(~ target_type,
+              scales = "free_x")
```

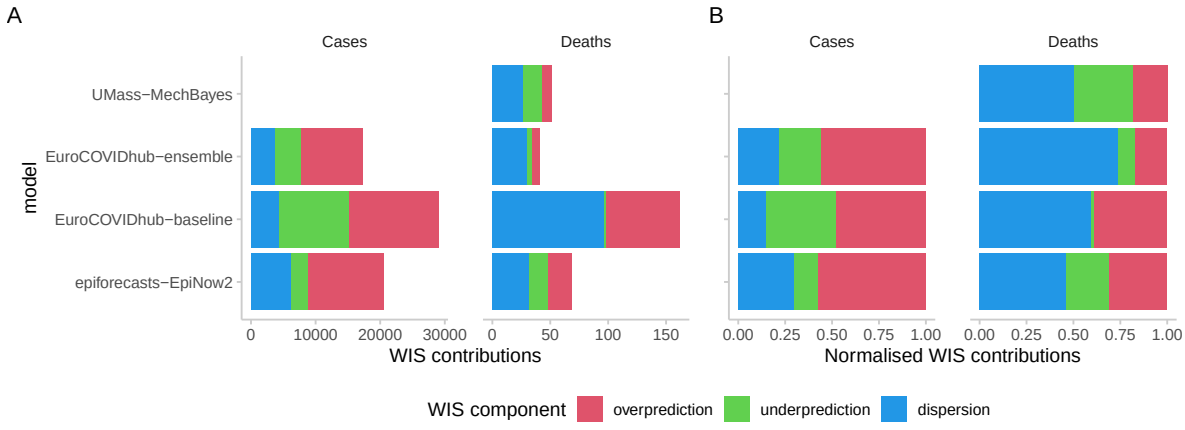

Figure 10: Decomposition of the weighted interval score (WIS) into dispersion, overprediction and underprediction. A: absolute contributions, B: contributions normalised to 1.

5. Discussion

Summary

This paper presented **scoringutils** an R package for forecast evaluation. It explained the core workflow, consisting of 1) validating and processing inputs, 2) scoring forecasts and 3) summarising scores, as well as additional functionality such as visualisation and diagnostic tools.

The package specialises in the evaluation of probabilistic forecasts (the forecast is a full predictive distribution). It provides a comprehensive framework based on **data.table** and allows users to validate, diagnose, visualise, transform and score forecasts using a wide range of default and custom scoring rules. The package is designed to be flexible and extensible, and to make it easy to use functionality from different packages in a single workflow. **scoringutils** addresses a gap in the existing ecosystem of forecast evaluation by creating a `data.table`-based forecast evaluation framework for probabilistic forecasts (similarly to what **yardstick** provides for point forecasts and classification tasks). Notably, **scoringutils** is the first package to provide extensive support for forecasts in a quantile-based forecasts, which is commonly used for example in Epidemiology. In addition to providing a coherent forecast evaluation workflow it offers a wide range of additional functions that practitioners may find useful when assessing or comparing the quality of their forecasts.

One important limitation of the package is that it currently does not support statistical testing of forecast performance as part of its core workflow. Determining whether a forecaster is significantly better than another is an important aspect of forecast evaluation that is currently mostly missing from the package. Another limitation is the fact that the package currently only supports a small set of possible types of forecasts. For example, forecasts in a bin-format or forecasts represented in a closed-form distribution (as can be scored for example using **scoringRules** are not supported. While it is in principle possible to extend the current classes and generic functions, this may not be very feasible in practice for most users. Some functionality in **scoringutils** is necessarily redundant with other packages that provide functionality to aid with the evaluation of forecasts. The overall idea of providing a `data.frame`-based evaluation framework, for example, is similar to what **yardstick** offers (albeit with a focus on point forecasts and classification tasks, rather than probabilistic forecasts). Having a single package that encompasses all possible use cases might be preferable. At the moment, **scoringutils** falls somewhat short of its aspiration to become a bridge between different packages in the forecast evaluation ecosystem. It does not yet offer a wide range of helper functions that allow users to easily convert between different formats and use functionality from other packages and many visualisations that are available in other packages, particularly with respect to model calibration, are missing.

A variety of extensions are planned for **scoringutils**. The first is the expansion of the forecast types that are supported. We plan to add support for evaluating categorical forecasts, as well as multivariate forecasts that specify a joint distribution across targets. Adding the possibility to score closed-form distributions might be another useful extension. A second area of expansion is the integration with other forecast evaluation and modelling packages. We aim to provide a variety of helper functions to convert to and from different formats, such as the one used by **yardstick** or formats used by modelling packages such as **odin**. These functions would make it easy to integrate **scoringutils** into existing workflows or use functionality from other packages that is not available in **scoringutils**. A third area of improvement is the addition of case studies and vignettes that make working with and extending functionality

from the package easier.

scoringutils is already used by a variety of public health institutions such as the US Centers for Disease Control, the European Centre for Disease Prevention and Control, as well as various academic institutions. The package is actively maintained and developed and we hope it will continue to be a valuable resource for researchers and practitioners working on forecast evaluation.

6. Acknowledgments

Funding statements

NIB received funding from the Health Protection Research Unit (grant code NIHR200908). HG MISSING. AC acknowledges funding by the NIHR, the Sergei Brin foundation, USAID, and the Academy of Medical Sciences. EvL acknowledges funding by the National Institute for Health Research (NIHR) Health Protection Research Unit (HPRU) in Modelling and Health Economics (grant number NIHR200908) and the European Union's Horizon 2020 research and innovation programme - project EpiPose (101003688). SF's work was supported by the Wellcome Trust (grant: 210758/Z/18/Z), and the NIHR (NIHR200908). SA's work was funded by the Wellcome Trust (grant: 210758/Z/18/Z). This study is partially funded by the National Institute for Health Research (NIHR) Health Protection Research Unit in Modelling and Health Economics, a partnership between UK Health Security Agency and Imperial College London in collaboration with LSHTM (grant code NIHR200908); and acknowledges funding from the MRC Centre for Global Infectious Disease Analysis (reference MR/R015600/1), jointly funded by the UK Medical Research Council (MRC) and the UK Foreign, Commonwealth & Development Office (FCDO), under the MRC/FCDO Concordat agreement and is also part of the EDCTP2 programme supported by the European Union. Disclaimer: "The views expressed are those of the author(s) and not necessarily those of the NIHR, UKHSA or the Department of Health and Social Care. We thank Community Jameel for Institute and research funding

References

Angus JE (1994). "The Probability Integral Transform and Related Results." *SIAM Review*, **36**(4), 652–654. ISSN 0036-1445. doi:10.1137/1036146.

Bosse NI, Abbott S, Cori A, van Leeuwen E, Bracher J, Funk S (2023). "Scoring Epidemiological Forecasts on Transformed Scales." *PLOS Computational Biology*, **19**(8), e1011393. ISSN 1553-7358. doi:10.1371/journal.pcbi.1011393.

Bracher J, Ray EL, Gneiting T, Reich NG (2021). "Evaluating Epidemic Forecasts in an Interval Format." *PLoS computational biology*, **17**(2), e1008618. ISSN 1553-7358. doi:10.1371/journal.pcbi.1008618.

Bracher J, Wolffram D, Deuschel J, Görgen K, Ketterer JL, Ullrich A, Abbott S, Barbarossa MV, Bertsimas D, Bhatia S, Bodych M, Bosse NI, Burgard JP, Castro L, Fairchild G, Fiedler J, Fuhrmann J, Funk S, Gambin A, Gogolewski K, Heyder S, Hotz T, Kheifetz

Y, Kirsten H, Krueger T, Krymova E, Leithäuser N, Li ML, Meinke JH, Miasojedow B, Michaud IJ, Mohring J, Nouvellet P, Nowosielski JM, Ozanski T, Radwan M, Rakowski F, Scholz M, Soni S, Srivastava A, Gneiting T, Schienle M (2022). "National and Subnational Short-Term Forecasting of COVID-19 in Germany and Poland during Early 2021." *Communications Medicine*, **2**(1), 1–17. ISSN 2730-664X. doi:10.1038/s43856-022-00191-8.

Cramer E, Ray EL, Lopez VK, Bracher J, Brennen A, Rivadeneira AJC, Gerding A, Gneiting T, House KH, Huang Y, Jayawardena D, Kanji AH, Khandelwal A, Le K, Mühlemann A, Niemi J, Shah A, Stark A, Wang Y, Wattanachit N, Zorn MW, Gu Y, Jain S, Bannur N, Deva A, Kulkarni M, Merugu S, Raval A, Shingi S, Tiwari A, White J, Woody S, Dahan M, Fox S, Gaither K, Lachmann M, Meyers LA, Scott JG, Tec M, Srivastava A, George GE, Cegan JC, Dettwiller ID, England WP, Farthing MW, Hunter RH, Lafferty B, Linkov I, Mayo ML, Parno MD, Rowland MA, Trump BD, Corsetti SM, Baer TM, Eisenberg MC, Falb K, Huang Y, Martin ET, McCauley E, Myers RL, Schwarz T, Sheldon D, Gibson GC, Yu R, Gao L, Ma Y, Wu D, Yan X, Jin X, Wang YX, Chen Y, Guo L, Zhao Y, Gu Q, Chen J, Wang L, Xu P, Zhang W, Zou D, Biegel H, Lega J, Snyder TL, Wilson DD, McConnell S, Walraven R, Shi Y, Ban X, Hong QJ, Kong S, Turtle JA, Ben-Nun M, Riley P, Riley S, Koyluoglu U, DesRoches D, Hamory B, Kyriakides C, Leis H, Milliken J, Moloney M, Morgan J, Ozcan G, Schrader C, Shakhnovich E, Siegel D, Spatz R, Stiefeling C, Wilkinson B, Wong A, Gao Z, Bian J, Cao W, Ferres JL, Li C, Liu TY, Xie X, Zhang S, Zheng S, Vespignani A, Chinazzi M, Davis JT, Mu K, y Piontti AP, Xiong X, Zheng A, Baek J, Farias V, Georgescu A, Levi R, Sinha D, Wilde J, Penna ND, Celi LA, Sundar S, Cavany S, España G, Moore S, Oidtman R, Perkins A, Osthus D, Castro L, Fairchild G, Michaud I, Karlen D, Lee EC, Dent J, Grantz KH, Kaminsky J, Kaminsky K, Keegan LT, Lauer SA, Lemaitre JC, Lessler J, Meredith HR, Perez-Saez J, Shah S, Smith CP, Truelove SA, Wills J, Kinsey M, Obrecht RF, Tallaksen K, Burant JC, Wang L, Gao L, Gu Z, Kim M, Li X, Wang G, Wang Y, Yu S, Reiner RC, Barber R, Gaikedu E, Hay S, Lim S, Murray C, Pigott D, Prakash BA, Adhikari B, Cui J, Rodríguez A, Tabassum A, Xie J, Keskinocak P, Asplund J, Baxter A, Oruc BE, Serban N, Arik SO, Dusenberry M, Epshteyn A, Kanal E, Le LT, Li CL, Pfister T, Sava D, Sinha R, Tsai T, Yoder N, Yoon J, Zhang L, Abbott S, Bosse NI, Funk S, Hellewel J, Meakin SR, Munday JD, Sherratt K, Zhou M, Kalantari R, Yamana TK, Pei S, Shaman J, Ayer T, Adee M, Chhatwal J, Dalgic OO, Ladd MA, Linas BP, Mueller P, Xiao J, Li ML, Bertsimas D, Lami OS, Soni S, Bouardi HT, Wang Y, Wang Q, Xie S, Zeng D, Green A, Bien J, Hu AJ, Jahja M, Narasimhan B, Rajanala S, Rumack A, Simon N, Tibshirani R, Tibshirani R, Ventura V, Wasserman L, O'Dea EB, Drake JM, Pagano R, Walker JW, Slayton RB, Johansson M, Biggerstaff M, Reich NG (2021). "Evaluation of Individual and Ensemble Probabilistic Forecasts of COVID-19 Mortality in the US." *medRxiv*, p. 2021.02.03.21250974. doi:10.1101/2021.02.03.21250974.

Cramer E, Reich NG, Wang SY, Niemi J, Hannan A, House K, Gu Y, Xie S, Horstman S, aniruddhadiga, Walraven R, starkari, Li ML, Gibson G, Castro L, Karlen D, Wattanachit N, jinghuichen, zyt9lsb, aagarwal1996, Woody S, Ray E, Xu FT, Biegel H, GuidoEspana, X X, Bracher J, Lee E, har96, leyouz (2020). "COVID-19 Forecast Hub: 4 December 2020 Snapshot." doi:10.5281/zenodo.3963371.

Czado C, Gneiting T, Held L (2009). "Predictive Model Assessment for Count Data." *Biometrics*, **65**(4), 1254–1261. ISSN 1541-0420. doi:10.1111/j.1541-0420.2009.01191.x.

Dawid AP (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), 278–290. ISSN 2397-2327. doi:10.2307/2981683.

Dowle M, Srinivasan A (2023). *data.table: Extension of 'data.frame'*. R package version 1.14.8, URL https://CRAN.R-project.org/package=data.table.

Elliott G, Timmermann A (2016). "Forecasting in Economics and Finance." *Annual Review of Economics*, **8**(1), 81–110. doi:10.1146/annurev-economics-080315-015346.

Epstein ES (1969). "A Scoring System for Probability Forecasts of Ranked Categories." *Journal of Applied Meteorology*, **8**(6), 985–987. ISSN 0021-8952. doi:10.1175/1520-0450(1969)008<0985:ASSFPF>2.0.CO;2.

Funk S, Abbott S, Atkins BD, Baguelin M, Baillie JK, Birrell P, Blake J, Bosse NI, Burton J, Carruthers J, Davies NG, Angelis DD, Dyson L, Edmunds WJ, Eggo RM, Ferguson NM, Gaythorpe K, Gorsich E, Guyver-Fletcher G, Hellewell J, Hill EM, Holmes A, House TA, Jewell C, Jit M, Jombart T, Joshi I, Keeling MJ, Kendall E, Knock ES, Kucharski AJ, Lythgoe KA, Meakin SR, Munday JD, Openshaw PJM, Overton CE, Pagani F, Pearson J, Perez-Guzman PN, Pellis L, Scarabel F, Semple MG, Sherratt K, Tang M, Tildesley MJ, van Leeuwen E, Whittles LK, Group CCW, Team ICCR, Investigators I (2020). "Short-Term Forecasts to Inform the Response to the Covid-19 Epidemic in the UK." *medRxiv*, p. 2020.11.11.20220962. doi:10.1101/2020.11.11.20220962.

Funk S, Camacho A, Kucharski AJ, Lowe R, Eggo RM, Edmunds WJ (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. ISSN 1553-7358. doi:10.1371/journal.pcbi.1006785.

Gneiting T, Balabdaoui F, Raftery AE (2007). "Probabilistic Forecasts, Calibration and Sharpness." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **69**(2), 243–268. ISSN 1467-9868. doi:10.1111/j.1467-9868.2007.00587.x.

Gneiting T, Raftery AE (2005). "Weather Forecasting with Ensemble Methods." *Science*, **310**(5746), 248–249. ISSN 0036-8075, 1095-9203. doi:10.1126/science.1115255.

Gneiting T, Raftery AE (2007). "Strictly Proper Scoring Rules, Prediction, and Estimation." *Journal of the American Statistical Association*, **102**(477), 359–378. ISSN 0162-1459, 1537-274X. doi:10.1198/016214506000001437.

Good IJ (1952). "Rational Decisions." *Journal of the Royal Statistical Society. Series B (Methodological)*, **14**(1), 107–114. ISSN 0035-9246. 2984087.

Hamill TM (2001). "Interpretation of Rank Histograms for Verifying Ensemble Forecasts." *Monthly Weather Review*, **129**(3), 550–560. ISSN 1520-0493, 0027-0644. doi:10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2.

Hamner B, Frasco M (2018). *Metrics: Evaluation Metrics for Machine Learning*. R package version 0.1.4, URL https://CRAN.R-project.org/package=Metrics.

Jordan A, Krüger F, Lerch S (2019). "Evaluating Probabilistic Forecasts with scoringRules." *Journal of Statistical Software*, **90**(12), 1–37. doi:10.18637/jss.v090.i12.

Kuhn M, Vaughan D, Hvitfeldt E (2023a). *yardstick: Tidy Characterizations of Model Performance*. R package version 1.2.0, URL https://CRAN.R-project.org/package=yardstick.

Kuhn M, Vaughan D, Ruiz E (2023b). *probably: Tools for Post-Processing Class Probability Estimates*. R package version 1.0.2, URL https://CRAN.R-project.org/package=probably.

Kuhn M, Wickham H (2020). *Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles*. URL https://www.tidymodels.org.

Kukkonen J, Olsson T, Schultz DM, Baklanov A, Klein T, Miranda AI, Monteiro A, Hirtl M, Tarvainen V, Boy M, Peuch VH, Poupkou A, Kioutsioukis I, Finardi S, Sofiev M, Sokhi R, Lehtinen KEJ, Karatzas K, San José R, Astitha M, Kallos G, Schaap M, Reimer E, Jakobs H, Eben K (2012). "A Review of Operational, Regional-Scale, Chemical Weather Forecasting Models in Europe." *Atmospheric Chemistry and Physics*, **12**(1), 1–87. ISSN 1680-7316. doi:10.5194/acp-12-1-2012.

Laboratory NRA (2015). *verification: Weather Forecast Verification Utilities*. R package version 1.42, URL https://CRAN.R-project.org/package=verification.

Liboschik T, Fokianos K, Fried R (2017). "tscount: An R Package for Analysis of Count Time Series Following Generalized Linear Models." *Journal of Statistical Software*, **82**(5), 1–51. doi:10.18637/jss.v082.i05.

Mann HB, Whitney DR (1947). "On a Test of Whether One of Two Random Variables Is Stochastically Larger than the Other." *The Annals of Mathematical Statistics*, **18**(1), 50–60. ISSN 0003-4851, 2168-8990. doi:10.1214/aoms/1177730491.

Matheson JE, Winkler RL (1976). "Scoring Rules for Continuous Probability Distributions." *Management Science*, **22**(10), 1087–1096. ISSN 0025-1909. doi:10.1287/mnsc.22.10.1087.

Merkle EC, Steyvers M (2013). "Choosing a Strictly Proper Scoring Rule." *Decision Analysis*, **10**, 292–304.

Meyer S, Held L, Höhle M (2017). "Spatio-Temporal Analysis of Epidemic Phenomena Using the R Package surveillance." *Journal of Statistical Software*, **77**(11), 1–55. doi:10.18637/jss.v077.i11.

Murphy AH (1971). "A Note on the Ranked Probability Score." *Journal of Applied Meteorology and Climatology*, **10**(1), 155–156. ISSN 1520-0450. doi:10.1175/1520-0450(1971)010<0155:ANOTRP>2.0.CO;2.

O'Hara-Wild M, Hyndman R, Wang E (2023). *fabletools: Core Tools for Packages in the 'fable' Framework*. R package version 0.3.4, URL https://CRAN.R-project.org/package=fabletools.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Reich NG, Brooks LC, Fox SJ, Kandula S, McGowan CJ, Moore E, Osthus D, Ray EL, Tushar A, Yamana TK, Biggerstaff M, Johansson MA, Rosenfeld R, Shaman J (2019). "A Collaborative Multiyear, Multimodel Assessment of Seasonal Influenza Forecasting in the United States." *Proceedings of the National Academy of Sciences*, **116**(8), 3146–3154. ISSN 0027-8424, 1091-6490. doi:10.1073/pnas.1812594116.

Rizopoulos D (2019). *cvGEE: Cross-Validated Predictions from GEE*. R package version 0.3-0, URL https://CRAN.R-project.org/package=cvGEE.

Rizopoulos D (2023). *GLMMadaptive: Generalized Linear Mixed Models using Adaptive Gaussian Quadrature*. R package version 0.9-0, URL https://CRAN.R-project.org/package=GLMMadaptive.

Sadatsafavi M, Safari A, Lee TY (2023). *predtools: Prediction Model Tools*. R package version 0.0.3, URL https://CRAN.R-project.org/package=predtools.

Sherratt K, Gruson H, Grah R, Johnson H, Niehus R, Prasse B, Sandman F, Deuschel J, Wolffram D, Abbott S, Ullrich A, Gibson G, Ray EL, Reich NG, Sheldon D, Wang Y, Wattanachit N, Wang L, Trnka J, Obozinski G, Sun T, Thanou D, Pottier L, Krymova E, Barbarossa MV, Leithäuser N, Mohring J, Schneider J, Wlazlo J, Fuhrmann J, Lange B, Rodiah I, Baccam P, Gurung H, Stage S, Suchoski B, Budzinski J, Walraven R, Villanueva I, Tucek V, Šmíd M, Zajícek M, Pérez ÁC, Reina B, Bosse NI, Meakin S, Di Loro A, Maruotti A, Eclerová V, Kraus A, Kraus D, Pribylova L, Dimitris B, Li ML, Saksham S, Dehning J, Mohr S, Priesemann V, Redlarski G, Bejar B, Ardenghi G, Parolini N, Ziarelli G, Bock W, Heyder S, Hotz T, E SD, Guzman-Merino M, Aznarte JL, Moriña D, Alonso S, Álvarez E, López D, Prats C, Burgard JP, Rodloff A, Zimmermann T, Kuhlmann A, Zibert J, Pennoni F, Divino F, Català M, Lovison G, Giudici P, Tarantino B, Bartolucci F, Jona LG, Mingione M, Farcomeni A, Srivastava A, Montero-Manso P, Adiga A, Hurt B, Lewis B, Marathe M, Porebski P, Venkatramanan S, Bartczuk R, Dreger F, Gambin A, Gogolewski K, Gruziel-Slomka M, Krupa B, Moszynski A, Niedzielewski K, Nowosielski J, Radwan M, Rakowski F, Semeniuk M, Szczurek E, Zielinski J, Kisielewski J, Pabjan B, Holger K, Kheifetz Y, Scholz M, Bodych M, Filinski M, Idzikowski R, Krueger T, Ozanski T, Bracher J, Funk S (2022). "Predictive Performance of Multi-Model Ensemble Forecasts of COVID-19 across European Nation." *Europe PMC*. doi:10.1101/2022.06.16.22276024.

Siegert S (2020). *SpecsVerification: Forecast Verification Routines for Ensemble Forecasts of Weather and Climate*. R package version 0.5-3, URL https://CRAN.R-project.org/package=SpecsVerification.

Timmermann A (2018). "Forecasting Methods in Finance." *Annual Review of Financial Economics*, **10**(1), 449–479. doi:10.1146/annurev-financial-110217-022713.

Yan Y (2016). *MLmetrics: Machine Learning Evaluation Metrics*. R package version 1.1.1, URL https://CRAN.R-project.org/package=MLmetrics.

Zeileis A, Lang MN (2022). *topmodels: Infrastructure for Inference and Forecasting in Probabilistic Models*. R package version 0.1-0/r1498, URL https://R-Forge.R-project.org/projects/topmodels/.

A. Constructing and validating forecast objects

The following section gives an overview of how **scoringutils** constructs forecast objects. The `forecast` class comes with a constructor, `new_forecast()`, a generic validation function, `validate_forecast()`, and a convenient wrapper function `as_forecast()`.

`new_forecast()` constructs a `forecast` object based on a `data.frame` or similar. It makes a deep copy of the inptut and converts it into a `data.table`, adds a `model` column with value "Unspecified model" if there isn't one and adds a class `forecast_*`, where `*` depends on the forecast type to the object.

`validate_forecast()` is a generic which dispatches to a specialised validator method depending on the class of the input. It validates the input and returns it if it is valid. If the input is not valid, it throws an error with a message that explains what went wrong.

`as_forecast()` (optionally) renames existing columns to conform with the requirements for forecast objects, (optionally) sets the forecast unit, determines the forecast type of the input (and optionally checks for consistency with what the user expects), constructs the class and validates the input. The process is illustrated in Figure A.11.

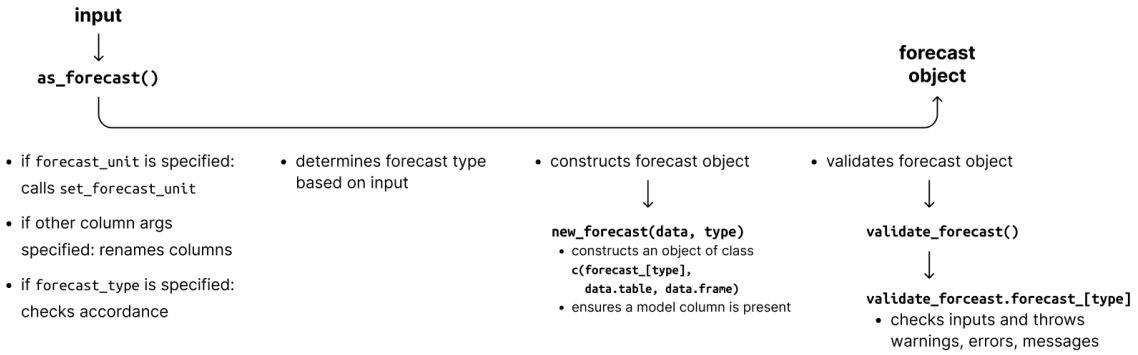

Figure A.11: Illustration of the process of creating a 'forecast' object.

B. Comparing different calibration plots

The following Figure gives a more detailed overview of how to interpret different calibration plots (showing the actual forecasts and observations that produced the corresponding visualisations).

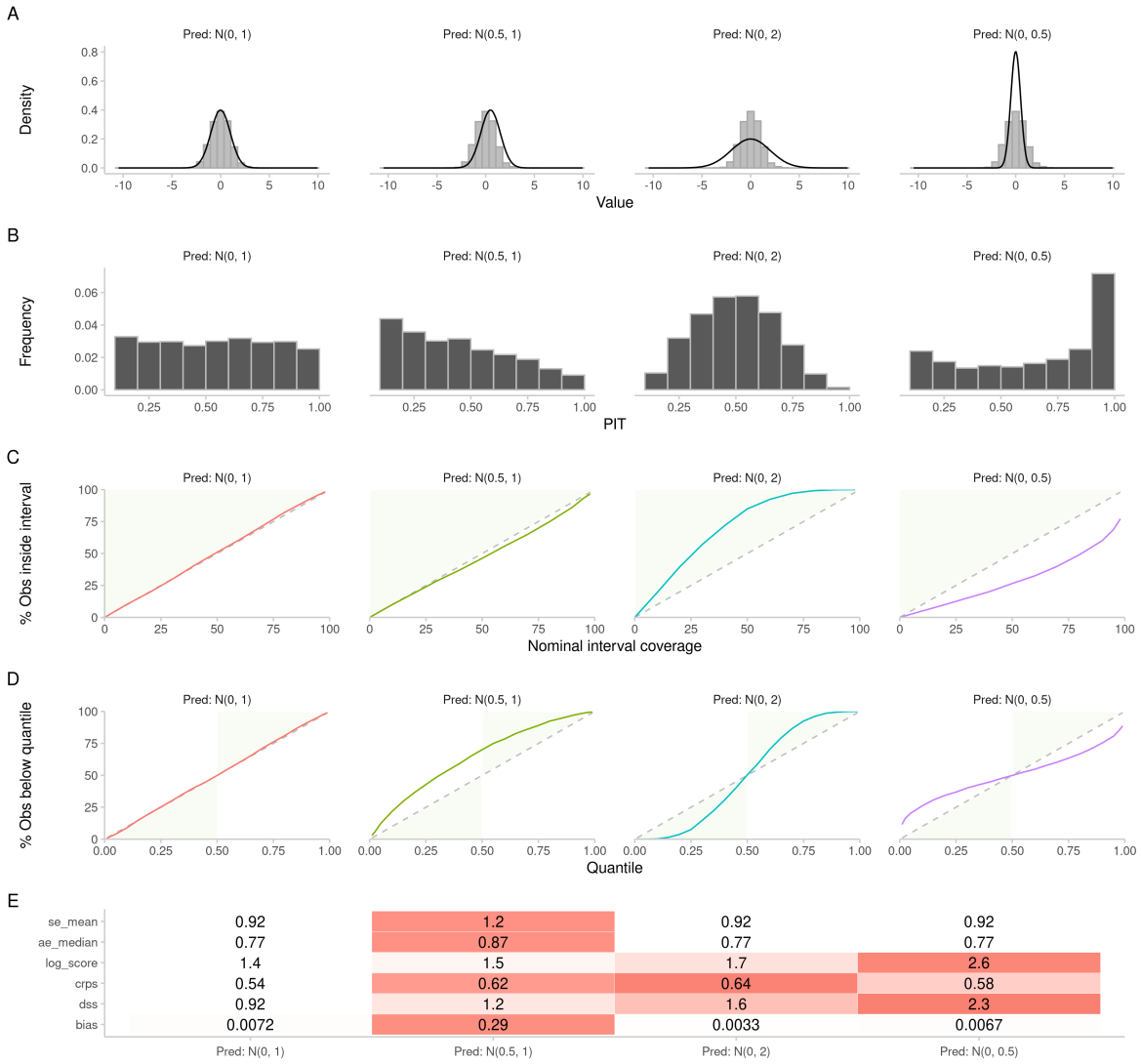

Figure B.12: A: Different forecasting distributions (black) against observations sampled from a standard normal distribution (grey histograms). B: PIT histograms based on the predictive distributions and the sampled observations shown in A. C: Empirical vs. nominal coverage of the central prediction intervals for simulated observations and predictions. Areas shaded in green indicate that the forecasts are too wide (i.e., underconfident), covering more true values than they actually should, while areas in white indicate that the model generates too narrow predictions and fails to cover the desired proportion of true values with its prediction intervals. D: Quantile coverage values, with green areas indicating too wide (i.e., conservative) forecasts. E: Scores for the standard normal predictive distribution and the observations drawn from different data-generating distributions.

C. Details on the weighted interval score (WIS)

The WIS treats the predictive quantiles as a set of symmetric prediction intervals and measures the distance between the observation and the forecast interval. It can be decomposed into a dispersion (uncertainty) component and penalties for over- and underprediction. For a single interval, the interval score is computed as

$$IS_\alpha(F, y) = \underbrace{(u - l)}_{\text{dispersion}} + \underbrace{\frac{2}{\alpha} \cdot (l - y) \cdot \mathbf{1}(y \leq l)}_{\text{overprediction}} + \underbrace{\frac{2}{\alpha} \cdot (y - u) \cdot \mathbf{1}(y \geq u)}_{\text{underprediction}},$$

where $\mathbf{1}()$ is the indicator function, y is the observed value, and l and u are the $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$ quantiles of the predictive distribution F, i.e. the lower and upper bound of a single prediction interval. For a set of K prediction intervals and the median m, the score is computed as a weighted sum,

$$WIS = \frac{1}{K + 0.5} \cdot \left(w_0 \cdot |y - m| + \sum_{k=1}^{K} w_k \cdot IS_\alpha(F, y) \right),$$

where w_k is a weight for every interval. Usually, $w_k = \frac{\alpha_k}{2}$ and $w_0 = 0.5$.

*Nikos I. Bosse, Hugo Gruson, Anne Cori, Edwin van Leeuwen, Sebastian Funk, Sam Abbott*33

Affiliation:

Nikos I. Bosse
London School of Hygiene & Tropical Medicine (LSHTM)
Centre for Mathematical Modelling of Infectious Diseases
London School of Hygiene & Tropical Medicine
Keppel Street
London WC1E 7HT
E-mail: nikos.bosse@lshtm.ac.uk
URL: https://lshtm.ac.uk

Hugo Gruson
LSHTM
Centre for Mathematical Modelling of Infectious Diseases
London School of Hygiene & Tropical Medicine
Keppel Street
London WC1E 7HT
E-mail: hugo.gruson@lshtm.ac.uk

Anne Cori
Imperial College London
MRC Centre for Global Infectious Disease Analysis, School of Public Health
Imperial College London
Norfolk Place
London W2 1PG
E-mail: a.cori@imperial.ac.uk

Edwin van Leeuwen
UK Health Security Agency, LSHTM
Statistics, Modelling and Economics Department
UK Health Security Agency
London NW9 5EQ
E-mail: Edwin.VanLeeuwen@phe.gov.uk

Sebastian Funk
LSHTM
Centre for Mathematical Modelling of Infectious Diseases
London School of Hygiene & Tropical Medicine
Keppel Street
London WC1E 7HT
E-mail: sebastian.funk@lshtm.ac.uk

Sam Abbott
LSHTM
Centre for Mathematical Modelling of Infectious Diseases
London School of Hygiene & Tropical Medicine
Keppel Street
London WC1E 7HT
E-mail: sam.abbott@lshtm.ac.uk

4 Comparing human and model-based forecasts of COVID-19 in Germany and Poland

This chapter investigates what human judgement can contribute to infectious disease forecasting, applying the tools and concepts covered in Chapters 2 and 3. The work in this chapter was motivated by the need to produce timely and useful forecasts of COVID-19. In October 2020 the German and Polish COVID0-19 Forecast Hub (Bracher et al., 2021b) launched, eliciting forecasts to help inform public health decision making in Germany and Poland. Previous submissions from our working group to the US COVID-19 Forecast Hub (Cramer et al., 2022) had proved cumbersome and it was not clear what the added benefit of mathematical modelling over human judgement alone was. We therefore developed an open source application, `crowdforecastr` (Bosse et al., 2020), which would allow humans to submit direct forecasts of cases and deaths in Germany and Poland. These forecasts were collected every week, aggregated and submitted to the German and Polish Forecast Hub. In order to obtain a better understanding of what mathematical modelling may add to human judgement alone, we submitted these forecasts alongside two mathematical models with minimal tuning. In an additional analysis, we further explore how adding different forecasts affects the quality of an ensemble of forecasts. This analysis was motivated by the desire to better understand what kinds of forecasts may contribute to public health decision making, and whether even imperfect forecast models can contribute usefully.

The study was limited both by resource constraints as well as the time that was available to develop the `crowdforecastr` platform, obtain ethics approval, conduct outreach, and the need to submit the first forecasts within 6 weeks of starting this PhD. The study should therefore better be understood as a case study that explores a variety of questions related to the interplay of human judgement and mathematical modelling, rather than providing definitive conclusions.

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Comparing human and model-based forecasts of COVID-19 in Germany and Poland

Nikos I. Bosse[1,2]*, Sam Abbott[1,2], Johannes Bracher[3], Habakuk Hain[4], Billy J. Quilty[1,2], Mark Jit[1,2], Centre for the Mathematical Modelling of Infectious Diseases COVID-19 Working Group[1,2], Edwin van Leeuwen[1,5], Anne Cori[6], Sebastian Funk[1,2]

1 Department of Infectious Disease Epidemiology, London School of Hygiene & Tropical Medicine, London, United Kingdom, **2** Centre for the Mathematical Modelling of Infectious Diseases (members of the CMMID COVID-19 working group are listed in S1 Acknowledgements), London, United Kingdom, **3** Institute of Economic Theory and Statistics, Karlsruhe Institute of Technology, Karlsruhe, Germany, **4** Max Planck Institute for Multidisciplinary Sciences, Göttingen, Germany, **5** UK Health Security Agency, London, United Kingdom, **6** MRC Centre for Outbreak Analysis and Modelling, Department of Infectious Disease Epidemiology, School of Public Health, Imperial College London, London, United Kingdom

* nikos.bosse@lshtm.ac.uk

OPEN ACCESS

Citation: Bosse NI, Abbott S, Bracher J, Hain H, Quilty BJ, Jit M, et al. (2022) Comparing human and model-based forecasts of COVID-19 in Germany and Poland. PLoS Comput Biol 18(9): e1010405. https://doi.org/10.1371/journal.pcbi.1010405

Editor: James M McCaw, The University of Melbourne, AUSTRALIA

Received: January 18, 2022

Accepted: July 18, 2022

Published: September 19, 2022

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcbi.1010405

Data Availability Statement: All code is available under https://github.com/epiforecasts/covid.german.forecasts.

Abstract

Forecasts based on epidemiological modelling have played an important role in shaping public policy throughout the COVID-19 pandemic. This modelling combines knowledge about infectious disease dynamics with the subjective opinion of the researcher who develops and refines the model and often also adjusts model outputs. Developing a forecast model is difficult, resource- and time-consuming. It is therefore worth asking what modelling is able to add beyond the subjective opinion of the researcher alone. To investigate this, we analysed different real-time forecasts of cases of and deaths from COVID-19 in Germany and Poland over a 1-4 week horizon submitted to the German and Polish Forecast Hub. We compared crowd forecasts elicited from researchers and volunteers, against a) forecasts from two semi-mechanistic models based on common epidemiological assumptions and b) the ensemble of all other models submitted to the Forecast Hub. We found crowd forecasts, despite being overconfident, to outperform all other methods across all forecast horizons when forecasting cases (weighted interval score relative to the Hub ensemble 2 weeks ahead: 0.89). Forecasts based on computational models performed comparably better when predicting deaths (rel. WIS 1.26), suggesting that epidemiological modelling and human judgement can complement each other in important ways.

Author Summary

Mathematical models of COVID-19 have played a key role in informing governments across the world. While mathematical models are informed by our knowledge of infectious disease dynamics, they are ultimately developed and iteratively adjusted by the researchers and shaped by their subjective opinions. To investigate what modelling is able to add beyond the subjective opinion of the researcher alone, we compared human forecasts with model-based predictions of COVID-19 cases and deaths submitted to the so-

Funding: NIB received funding from the National Institute for Health Research (NIHR) Health Protection Research Unit (grant code NIHR200908, https://www.nihr.ac.uk/). SA's work was funded by the Wellcome Trust (grant: 210758/Z/18/Z, https://wellcome.org/). The work of JB was supported by the Helmholtz Foundation (https://www.helmholtz.de/) via the SIMCARD Information and Data Science Pilot Project. The work of BJQ was partly funded by the National Institute for Health Research (NIHR, https://www.nihr.ac.uk/) (16/137/109 & 16/136/46) using UK aid from the UK Government to support global health research. The views expressed in this publication are those of the author(s) and not necessarily those of the NIHR or the UK Department of Health and Social Care. BJQ is also supported in part by a grant from the Bill and Melinda Gates Foundation (OPP1139859, https://www.gatesfoundation.org/). MJ and EvL acknowledge funding by the National Institute for Health Research (NIHR) Health Protection Research Unit (HPRU) in Modelling and Health Economics (grant number NIHR200908, https://www.nihr.ac.uk/) and the European Union's Horizon 2020 research and innovation programme - project EpiPose (101003688, https://ec.europa.eu/programmes/horizon2020/). AC acknowledges funding by the NIHR, the Sergei Brin foundation, USAID (https://www.usaid.gov/), and the Academy of Medical Sciences (https://acmedsci.ac.uk/). SF's work was supported by the Wellcome Trust (grant: 210758/Z/18/Z, https://wellcome.org/). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

called German/Polish Forecast Hub (which collates a variety of models from a range of teams). We found that our human forecasts consistently outperformed an aggregate of all available model-based forecasts when predicting cases, but not when predicting deaths. Our findings suggest that human insight may be most valuable when forecasting highly uncertain quantities, which depend on many factors that are hard to model using equations, while mathematical models may be most useful in settings like predicting deaths, where leading indicators with a clear connection to the target variable are available. This potentially has very relevant policy implications, as agencies informing policy-makers could benefit from routinely eliciting human forecasts in addition to model-based predictions to inform policies.

Introduction

Infectious disease modelling has a long tradition and has helped inform public health decisions both through scenario modelling, as well as actual forecasts of (among others) influenza [e.g. 1,2–4], dengue fever [e.g. 5,6,7], ebola [e.g. 8,9], chikungunya [e.g. 10,11] and now COVID-19 [e.g. 12,13–17]. Applications of epidemiological models differ in the way they make statements about the future. Forecasts aim to predict the future as it will occur, while scenario modelling and projections aim to represent what the future could look like under certain scenario assumptions or if conditions stayed the same as they were in the past. Forecasts can be judged by comparing them against observed data. Since it is much harder to fairly assess the accuracy and usefulness of projections and scenario modelling in the same way, this work focuses on forecasts, which represent only a subset of all epidemiological modelling.

Since March 2020, forecasts of COVID-19 from multiple teams have been collected, aggregated and compared by Forecast Hubs such as the US Forecast Hub [13, 14], the German and Polish Forecast Hub [15, 16] and the European Forecast Hub [17]. Often, different individual forecasts are combined into a single forecast, e.g. by taking the mean or median of all forecasts. These ensemble forecasts usually tend to perform better and more consistently than individual forecasts (see e.g. [6]; [18]).

Individual computational models usually rely to varying degrees on mechanistic assumptions about infectious disease dynamics (such as SIR-type compartmental models that aim to represent how individuals move from being susceptible to infected and then recovered or dead). Some are more statistical in nature (such as time series models that detect statistical patterns without explicitly modelling disease dynamics). How exactly such a mathematical or computational model is constructed and which assumptions are made depends on subjective opinion and judgement of the researcher who develops and refines the model. Models are commonly adjusted and improved based on whether the model output looks plausible to the researchers involved.

The process of model construction and refinement is laborious and time-consuming, and it is therefore worth asking what modelling can add beyond the subjective judgment of the researcher alone. In this work, we ask this question specifically in the context of predictive performance, and set aside other advantages of epidemiological modelling (such as reproducibility or the ability to obtain a deeper fundamental understanding of how diseases spread). One natural way to do this is to compare the predictive performance of forecasts based on computational models ("model-based forecasts") against forecasts made by individual humans without explicit use of a computer model ("direct human forecasts") or a combination of multiple such forecasts ("crowd forecasts").

Previous work has examined such direct human forecasts in various contexts, such as geopolitics [19, 20], meta-science [21, 22], sports [23] and epidemiology [11, 24, 25]. Several prediction platforms [26–28] and prediction markets [29] have been created to collate expert and non-expert predictions. However, with the notable exception of [11], these forecasts were not designed to be evaluated alongside model-based forecasts and usually follow their own (often binary) prediction formats. Direct human forecasts may be able to take into account insights and relationships between variables which are hard to specify using epidemiological models. However, it is not entirely clear in which situations human forecasts perform well or badly. For example, [11] found that humans could outperform computer models at predicting the 2014/15 and 2015/16 flu season in the US, a setting where the disease was well known and information about previous seasons was available. However, humans tended to do slightly worse at predicting the 2014/15 outbreak of chikungunya in the Americas, a disease previously largely unobserved and unknown in these regions at the time.

In this study, we analyse the performance of direct human forecasts relative to model-based forecasts and discuss the added benefit of epidemiological modelling over human judgement alone. As a case study, we use different forecasts, involving varying degrees of human intervention, which we submitted in real time to the German and Polish Forecast Hub. In contrast to [11] we elicited not only point predictions, but full predictive distributions ("probabilistic forecasts", see e.g. [30]) from participants. This allows us to compare not only predictive accuracy, but also how well human forecasters and model-based forecasts were able to quantify forecast uncertainty.

Methods

Ethics statement

This study has been approved by the London School of Hygiene & Tropical Medicine Research Ethics Committee (reference number 22290). Consent from participants was obtained in written form.

Overview

We created and submitted the following forecasts to the German and Polish Forecast Hub: 1) a direct human forecast (henceforth called "crowd forecast"), elicited from participants through a web application [31] and 2) two semi-mechanistic model-based forecasts ("renewal model" and "convolution model") informed by basic assumptions about COVID-19 epidemiology. While the two semi-mechanistic forecasts were necessarily shaped by our implicit assumptions and decisions, they were designed such as to minimise the amount of human intervention involved. For example, we refrained from adjusting model outputs or refining the models based on past performance. Forecasts were created in real time over a period of 21 weeks from October 12th 2020 until March 1st 2021 and submitted to the German and Polish Forecast hub [15, 16]. All code and tools necessary to generate the forecasts and make a forecast submission are available in the `covid.german.forecasts` R package [32]. This repository also contains a record of all forecasts submitted to the German and Polish Forecast Hub. Forecasts were evaluated using a variety of scoring metrics and compared among each other and against an ensemble of all other models submitted to the German and Polish Forecast Hub.

Forecast targets and interaction with the German and Polish Forecast Hub

The German and Polish Forecast Hub (now mostly merged into the European Forecast Hub [17]) elicits predictions for various COVID-19 related forecast targets from different research

groups every week. Forecasts had to be made every Monday (with submissions allowed until Tuesday 3pm) and were permitted to use any data that was available by Monday 11.59pm. We submitted forecasts for incident and cumulative weekly reported numbers of cases of and deaths from COVID-19 on a national level in Germany and Poland over a one to four week forecast horizon. Forecasts were submitted on Mondays, but weeks were defined as ending on a Saturday (and starting on Sunday), meaning that forecast horizons were in fact 5, 12, 19 and 26 days. Submissions were required in a quantile-based format with 23 quantiles of each output measure at levels 0.01, 0.025, 0.05, 0.10, 0.15,. . ., 0.95, 0.975, 0.99. Forecasts submitted to the Forecast Hub were combined into different ensembles every week, with the median ensemble (i.e., the α-quantile of the ensemble is given by the median of all submitted α-quantiles) being the default ensemble shown on all official Forecast hub visualisations (https://kitmetricslab.github.io/forecasthub/forecast).

Data on daily reported test positive cases and deaths linked to COVID-19 were provided by the organisers of the German and Polish Forecast hub. Until December 14th, 2020, these data were sourced from the European Centre for Disease Control [33]. After ECDC stopped publishing daily data, observations were sourced from the Robert Koch Institute (RKI) and the Polish Ministry of Health for the remainder of the submission period [34]. These data are subject to reporting artefacts, (such as for example delayed case reporting in Poland on the 24th November, [35]), changes in reporting over time, and variation in testing regimes (for example in Germany from the 11th of November on, [36]). The ECDC data as well as the data published by the Polish Ministry of Health were also subject to data revisions, although most of them (with a notable exception of a data update for October 12 2020 in Germany) only affected daily, not weekly data (see S7 and S8 Figs).

Crowd forecasts

Our crowd forecasts were created as an ensemble of forecasts made by individual participants every week through a web application (https://cmmid-lshtm.shinyapps.io/crowd-forecast/). Weekly forecasts had to be submitted before Tuesday 12pm every week, but participants were asked to only use any information or data that was already available by Monday night. The application was built using the `shiny` and `golem` R packages [37, 38] and is available in the `crowdforecastr` R package [31]. To make a forecast in the application participants could select a predictive distribution (with the default being log-normal) to represent the probability that the forecasted quantity took certain values. Median and width of the uncertainty could be adjusted by either interacting with a figure showing their forecast or providing numerical values (see screenshot in S1 Fig). The default shown was a repetition of the last known observation with constant uncertainty around it computed as the standard deviation of the last four changes in weekly log observed forecasts (i.e. as $\sigma(log(value4) - log(value3), log(value3) - log(value2), \ldots)$). A comparison of the crowd forecasts against the default baseline shown in the application is displayed in S25 Fig. Our interface also allowed participants to view past observations based on the hub data, as well as their forecasts, on a logarithmic scale and presented additional contextual COVID-19 data sourced from [39]. These data included, for example, notifications of both test positive COVID-19 cases and COVID-19 linked deaths and the number of COVID-19 tests conducted over time. From November 26 2020 on we displayed weekly small reports with a visualisation of past forecasts and scores on our website, epiforecasts.io.

Forecasts were stored in a Google Sheet and downloaded, cleaned and processed every week for submission to the Forecast Hub. If a forecaster had submitted multiple predictions for a single target, only the latest submission was kept. Information on the chosen distribution as well as the parameters for median and width were used to obtain the required set of 23

quantiles from that distribution. Forecasts from all forecasters were then aggregated using an unweighted quantile-wise mean (i.e., the α-quantile of the ensemble is given by the mean of all submitted α-quantiles). To avoid issues with users trying out the app and submitting a random forecast, we required that a forecaster needed to make a forecast for at least two targets for a given forecast in order to be included in the crowd forecast ensemble. On a few occasions we deleted forecasts that were clearly the result of a user or software error (such as for example forecasts that were zero everywhere).

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. Depending on whether they had a background in either statistics, forecasting or epidemiology, participants were asked to self-identify as 'experts' or 'non-experts'.

Model-based forecasts

We used two Bayesian semi-mechanistic models from the `EpiNow2` R package (version 1.3.3) as our model-based forecasts [40]. The first of these models, here called "renewal model", used the renewal equation [41] to predict reported cases and deaths (see details in S1 Text). It estimated the effective reproduction number R_t (the average number of people each person infected at time t is expected to infect in turn) and modelled future infections as a weighted sum of past infection multiplied by R_t. R_t was assumed to stay constant beyond the forecast date, roughly corresponding to continuing the latest exponential trend in infections. On the 9th of November we altered the date when R_t was assumed to be constant from two weeks prior to the date of the forecast to the forecast date, which we found to yield a more stable R_t estimate. Reported case and death notifications were obtained by convolving predicted infections over data-based delay distributions [40, 42–44] to model the time between infection and report date. The renewal model was used to predict cases as well as deaths with forecasts being generated for each target separately. Death forecasts from the renewal model were therefore not informed by past cases. One submission of the renewal model on December 28th 2020 was delayed and therefore not included in the official Forecast hub ensemble.

The second model ("convolution model", see details in S1 Text). was only used to forecast deaths and was added later, starting December 7th 2020 (with the first forecast from December 7th suffering from a software bug and therefore disregarded in all further analyses). The convolution model was submitted, but never included in the official Forecast hub ensemble due to concerns that it could be too similar to the renewal model. The convolution model predicted deaths as a fraction of infected people who would die with some delay, by using a convolution of reported cases with a distribution that described the delay from case report to death and a scaling factor (the case-fatality ratio). Both the renewal and the convolution model used daily observations and assumed a negative binomial observation model with a multiplicative day-of-the-week effect [40].

Line list data used to inform the prior for the delay from symptom onset to test positive case report or death in the model-based forecasts was sourced from [45] with data available up to the 1st of August. All model fitting was done using Markov-chain Monte Carlo (MCMC) in stan [46] with each location and forecast target being fitted separately.

Analysis

For the main analysis we focused mostly on two week ahead forecasts, as COVID-19 forecasts, especially for cases, were in the past found to have poor predictive performance beyond this horizon [15]. Forecasts for cases were scored using the full period from October 2020 until

March 2021. To ensure comparability between models, all death forecasts were scored using only the period from December 14th on, where all models including the convolution model were available. To ensure robustness of our results we conducted a sensitivity analysis where all forecasts (including cases) were scored only over the later period for which all forecasts were available (see S22 Fig and S8 and S9 Tables). Results remained broadly unchanged.

Forecasts were analysed using the following scoring metrics: The weighted interval score (WIS) [47], the absolute error, relative bias, and empirical coverage of the 50% and 90% prediction intervals. The WIS is a proper scoring rule [48], meaning that in expectation the score is optimised by reporting a predictive distribution that is identical to the true data-generating distribution. Forecasters are therefore incentivised to report their true belief about the future. The WIS can be understood as a generalisation of the absolute error to quantile-based forecasts (also meaning that smaller values are better) and can be decomposed into three separate penalties: forecast spread (i.e. uncertainty of forecasts), over-prediction and under-prediction. While the over- and under-prediction components of the WIS capture the amount of over-prediction and under-prediction in absolute terms, we also look at a relative tendency to make biased forecasts. The bias metric [9] we use captures how much probability mass of the forecast was above or below the true value (mapped to values between -1 and 1) and therefore represents a general tendency to over- or under-predict in relative terms. A value of -1 implies that all quantiles of the predictive distribution are below the observed value and a value of 1 that all quantiles are above the observed value. Empirical coverage is the percentage of observed values that fall inside a given prediction interval (e.g. how many observed values fall inside all 50% prediction intervals). Scoring metrics are explained in more detail in S1 Table. All scores were calculated using the `scoringutils` R package [49].

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"). This "Hub ensemble" was retrospectively computed and excludes all our models, leaving on average five ensemble member models (see S10 Table and S24 Fig). What we call "Hub ensemble" in this article therefore differs from the "official Hub ensemble" (here called "hub-ensemble-realised") which included crowd forecasts as well as renewal model forecasts. To enhance interpretability of scores we mainly report WIS relative to the Hub ensemble in the main text, i.e. we divided the average scores for a given model by the average score achieved by the Hub ensemble on the same set of forecasts (with values >1 implying worse and values <1 implying better performance than the Hub ensemble). In addition to comparing our forecasts against the hub ensemble excluding our models, we also assessed the impact of our forecasts on the performance of the forecasting hub by recalculating separate versions of the Hub ensemble with only some (or all) of our forecasts included. Versions that included either all of our models ("hub-ensemble-with-all") or only one of them ("hub-ensemble-with-X") were computed retrospectively.

Results

Crowd forecast participation

A total number of 32 participants submitted forecasts, 17 of those self-identified as 'expert' in either forecasting or epidemiology. The median number of forecasters for any given forecast target was 6, the minimum 2 and the maximum 10. 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 of whom are authors on this study.

Case forecasts

For cases, crowd forecasts had a lower mean weighted interval score (WIS, lower values indicate better performance) than both the renewal model and the Hub ensemble across all forecast horizons (Fig 1A) and locations (S5(A) Fig). For two week ahead forecasts, mean WIS relative to the Hub ensemble (= 1) was 0.89 for crowd forecasts and 1.40 for the renewal model (S2 Table). Across all forecasting approaches, locations and forecast horizons, the distribution of WIS values was very right-skewed, and average performance was heavily influenced by outliers (see Fig 2). Overall, low variance in forecast performance was closely linked with good

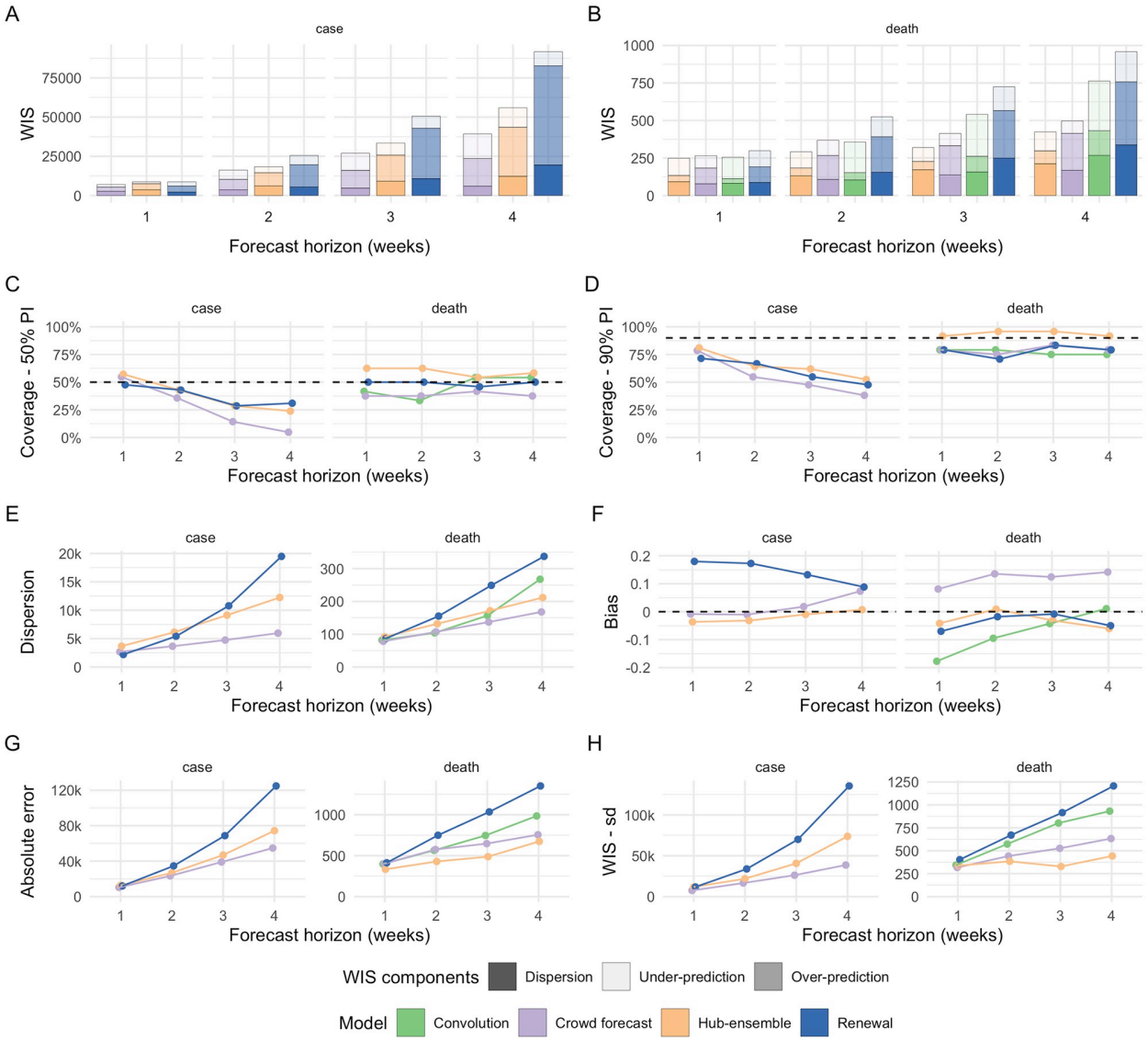

Fig 1. Visualisation of aggregate performance metrics for forecasts one to four weeks into the future. A, B: mean weighted interval score (WIS, lower indicates better performance) across horizons. WIS is decomposed into its components dispersion, over-prediction and under-prediction. C: Empirical coverage of the 50% prediction intervals (50% coverage is perfect). D: Empirical coverage of the 90% prediction intervals. E: Dispersion (same as in panel A, B). Higher values mean greater dispersion of the forecast and imply ceteris paribus a worse score. F: Bias, i.e. general (relative) tendency to over- or underpredict. Values are between -1 (complete under-prediction) and 1 (complete over-prediction) and 0 ideally. G: Absolute error of the median forecast (lower is better). H. Standard deviation of all WIS values for different horizons.

https://doi.org/10.1371/journal.pcbi.1010405.g001

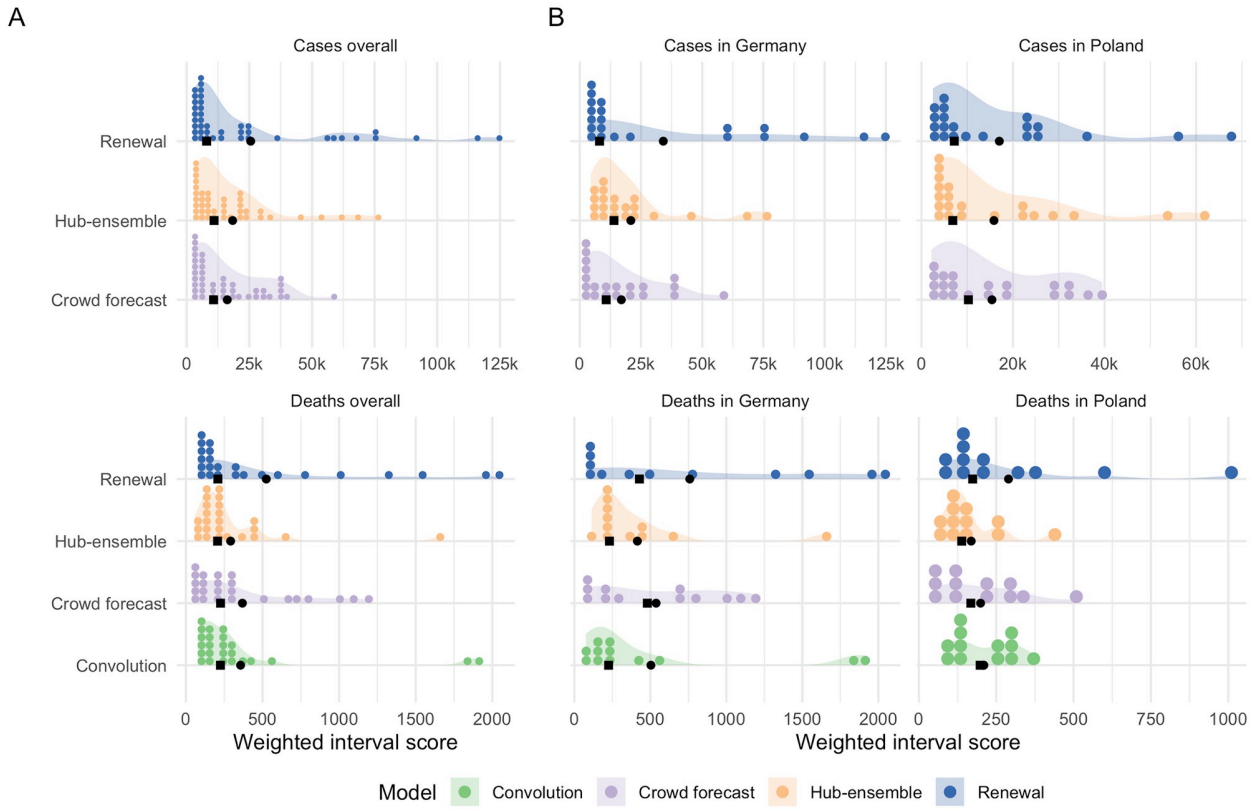

Fig 2. Two week ahead forecasts and corresponding scores. A, C: Visualisation of 50% prediction intervals of two week ahead forecasts against the reported values. Forecasts that were not scored (because there was no complete set of death forecasts available) are greyed out. B, D: Visualisation of corresponding WIS.

https://doi.org/10.1371/journal.pcbi.1010405.g002

mean performance (Fig 1H and 1A), suggesting that the ability to avoid large errors was an important factor in determining overall performance. The impact of outlier values was especially pronounced for the renewal model, which had more outliers, as well as the highest standard deviation of WIS values (standard deviation of the WIS relative to the WIS sd of the Hub ensemble was 1.54 at the two weeks ahead horizon), while the ensemble of crowd forecasts (rel. WIS sd 0.76) and the Hub ensemble (= 1) showed more stable performance.

To varying degrees, all forecasts exhibited trend-following behaviour and were rarely able to predict a change in trend before it had happened. For example, all forecasts failed to predict the change in trend from increase to decrease that happened in November in Germany and severely overshot reported cases (Fig 3A). This was most striking for the renewal model, which extrapolated unconstrained exponential growth based on the recent past of observations. The Hub ensemble and the crowd forecast, which had both been under-predicting throughout October, also failed to predict the change in trend after cases peaked, but less severely so. Human forecasters, possibly aware of the semi-lockdown announced on November 2nd 2020 [50] and the change in the testing regime (with stricter test criteria) on November 11th 2020 [36], were fastest to adapt to the new trend, and the Hub ensemble slowest. In December, cases rose again in Germany, with all models under-predicting this growth to varying extents. As in October, the renewal model captured the phase of exponential growth in cases slightly better than other approaches, but again overshot when reported case numbers fell over Christmas.

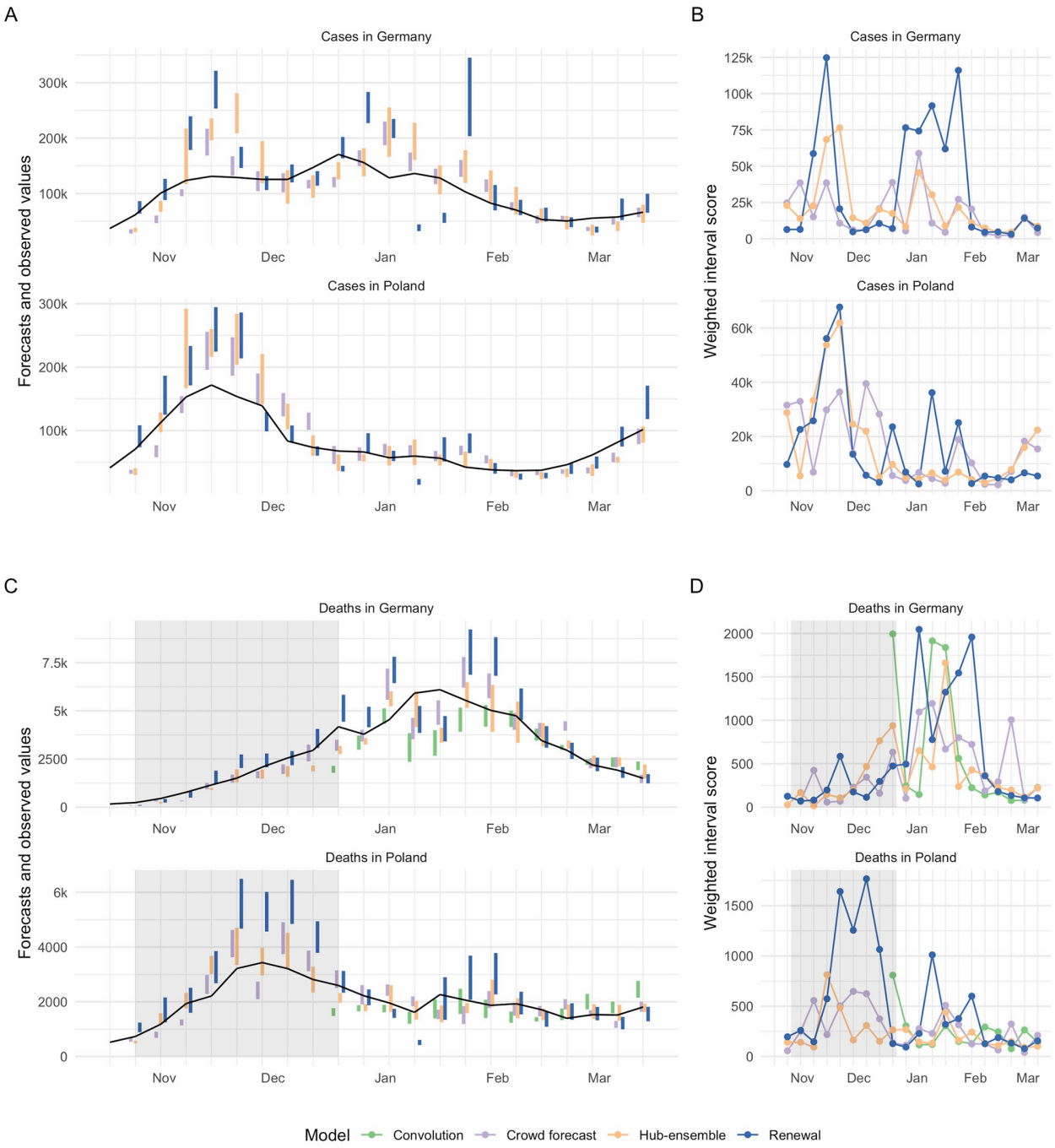

Fig 3. Distribution of scores. A: Distribution of weighted interval scores for two week ahead forecasts of the different models and forecast targets. Points denote single forecasts scores, while the shaded area shows an estimated probability density. B: Distribution of WIS separate by country. Black squares indicate median and black circles mean scores.

https://doi.org/10.1371/journal.pcbi.1010405.g003

The large variance in predictions in January in Germany (severe under-prediction followed by severe over-prediction) may in part be caused by the fact that the renewal model operated on daily data and therefore was susceptible to fluctuations in daily reporting around Christmas that would not have influenced on weekly reporting. Similar trends in performance were

evident in Poland, with the crowd forecast quickest at adapting to the change in trend in November. In general, there were fewer large outlier forecasts in Poland and in particular the renewal model performed more in line with other forecasts there.

All forecasting approaches, including the Hub ensemble, were overconfident, i.e. they showed lower than nominal coverage (meaning that 50% (90%) prediction intervals generally covered less than 50% (90%) of the actually observed values) (Fig 1C and 1D). Coverage for all forecasts deteriorated with increasing forecast horizon, indicating that all forecasting approaches struggled to quantify uncertainty appropriately for case forecasts. This was especially an issue for crowd forecasts, which had markedly shorter prediction intervals (i.e., narrower and more confident predictive distributions) than other approaches (Fig 1E) and only showed a small increase in uncertainty across forecast horizons. The crowd forecasts prediction intervals were also noticeably narrower than the default baseline shown to forecasters in the application (see S25 Fig).

In spite of good performance in terms of the absolute error (Fig 1G), the narrow forecast intervals led to forecasts which were severely overconfident (covering only 36% and 55% of all observations with the 50% and 90% prediction intervals of all forecasts made at a two week forecast horizon, and only 5% and 38% four weeks ahead) (Fig 1C and 1D as well as S2 and S3 Tables). Despite worse performance in terms of absolute error (Fig 1G), the renewal model achieved better calibration (comparable to the Hub ensemble), as uncertainty increased rapidly across forecast horizons. The crowd forecasts, on the other hand, showed a smaller bias than the renewal model, but were overconfident.

The renewal model exhibited a noticeable tendency towards over-predicting reported cases across all horizons. The crowd forecast tended to over-predict at longer forecast horizons, whereas the Hub ensemble showed no systematic bias (Fig 1F). Regardless of a general relative tendency to over-predict, all forecasting approaches incurred larger absolute penalties from over- than from under-prediction (see decomposition of the WIS into absolute penalties for over-prediction, under-prediction and dispersion in Fig 1A and 1B, as well as S2 and S3 Tables).

Generally, trends in overall performance were broadly similar across locations (S4 and S5 Figs). Due to the differing population sizes and numbers of notifications in Germany and Poland absolute scores were difficult to compare directly. However, relative to the Hub ensemble, the crowd forecasts performed noticeably better in Germany than in Poland and the renewal model better in Poland than in Germany (S5(A), S5(G), S2 and S3 Figs).

Death forecasts

For deaths, the Hub ensemble outperformed the crowd forecasts as well as our model-based approaches across all forecast horizons and locations (Fig 1B and S4(B) Fig). Relative WIS values for the models two weeks ahead were 1.22 (convolution model), 1.26 (crowd forecast), 1 (Hub ensemble) and 1.79 (renewal model). The crowd forecasts performed better than the renewal model across all forecast horizons and locations (Fig 1B and S4(B) Fig), and also better than the convolution model three and four weeks ahead. Poor performance of the renewal model, especially at longer horizons, indicates that an approach that does not know about past cases, but instead estimates and projects a separate R_t trace from deaths, does not use the available information efficiently. The convolution model was able to outperform both the renewal model and the crowd forecasts at shorter forecast horizons (where the delay between cases and deaths means that future deaths are largely informed by present cases), but saw performance deteriorate at three and four weeks ahead (where case predictions from the renewal model were increasingly used to inform death predictions) (Fig 1B, S3 Table).

As past cases and hospitalisations can be used as predictors, predicting a change in trend may be easier for deaths than for cases. Even though all forecasts generally struggled with this, there were some instances where changing trends were well captured or even anticipated. In Poland, for example, the Hub ensemble was able to capture or even anticipate the peak in deaths in December quite well (whereas the renewal model and crowd forecast did not). The renewal model, which mostly exhibited trend-following behaviour, correctly predicted another increase in weekly deaths in mid-January (potentially based on changes in daily deaths, as the renewal model did not know about past cases). In Germany in early January, all models predicted a decrease in deaths two to three weeks before it actually happened. Predictions from the renewal model at that time were likely strongly influenced by an unexpected drop in reported deaths in December. The other forecasting approaches and in particular, the convolution model may have been affected by potentially under-reported case numbers around Christmas. When the decrease that all models had predicted to happen in early January failed to materialise, the renewal model and the crowd forecast noticeably over-corrected and over-predicted deaths in the following weeks, while the Hub ensemble, and to a slightly lesser degree, the convolution model were able to capture the downturn well when it finally happened at the end of January.

Death forecasts, generally, showed greater coverage of the 50% and 90% prediction intervals than case forecasts and no decrease in coverage across forecast horizons, indicating that it might be easier to appropriately quantify uncertainty for death forecasts. The Hub ensemble had the greatest coverage, with empirical coverage of the 50% and 90% prediction intervals exceeding 50%, and 90%, respectively, across all forecast horizons. Coverage for the crowd forecasts and our model-based approaches was generally lower than that of the Hub ensemble and mostly slightly lower than nominal coverage (Fig 1C and 1D). As for cases, the crowd forecast tended to have the narrowest prediction intervals and uncertainty increased most slowly across forecast horizons, and the renewal model forecasts generally were widest. The convolution model had relatively narrow prediction intervals for short forecast horizons, but had rapidly (and non-linearly) increasing uncertainty for longer forecast horizons, driven by increasing uncertainty in the underlying case forecasts.

For deaths, the ensemble of crowd forecasts had a consistent tendency to over-predict (see Fig 1F). The convolution model had a strong tendency to under-predict, with the magnitude of under-prediction steadily decreasing for longer forecast horizons. The renewal model (which over-predicted for cases) and the Hub ensemble slightly tended towards under-prediction. For deaths, absolute over- and under-prediction penalties were more in line with a general relative tendency to over- or under-predict than for cases (Fig 1A and 1B, as well as S2 and S3 Tables).

Contribution to the forecast Hub

Of our three models, only the renewal model and the crowd forecast were included in the official Forecast Hub median ensemble ("hub-ensemble-realised"), while the convolution model was never included as it was deemed too similar to the existing renewal model. In the official Hub ensemble, there were on average 7.1 models included (including our own), with a median of 7, a minimum of 4 (on 28 December 2020 over the Christmas period) and a maximum of 10. Versions that included either all of our models ("hub-ensemble-with-all") or only one of them ("hub-ensemble-with-X") were computed retrospectively. An overview of all models and ensemble versions is shown in S10 Table.

For cases, our contributions (compared to the Hub ensemble without our contributions) consistently improved performance across all forecasting horizons (rel. WIS 0.9 two weeks

ahead, see S4 Table). Contributions from the crowd forecasts alone also improved performance of the Hub ensemble across all forecast horizons, while contributions from the renewal model had a negative effect for longer horizons (rel. WIS 1.02 three weeks ahead, 1.06 four weeks ahead). The realised ensemble including both models performed better or equal compared to all versions with only one model included for up to three weeks ahead, suggesting synergistic effects. Only for predictions four weeks ahead would removing the renewal model have improved performance (S5 Table). The realised ensemble performed comparably to the crowd forecasts for predictions one to two weeks ahead, and worse for greater forecast horizons.

For deaths, contributions from the renewal model and crowd forecast together improved performance only for one week ahead predictions and showed an increasingly negative impact on performance for longer horizons (rel. WIS of the Hub-ensemble-realised 1.01 two weeks ahead, 1.05 four weeks ahead, S4 and S5 Tables). Individual contributions from both the renewal model and the crowd forecast were largely negative, while a version of the Hub ensemble with only the convolution model included would have performed consistently better across all forecast horizons (with the positive impact increasing for longer horizons). This is especially interesting as the convolution model performed consistently worse than the pre-existing Hub ensemble (Fig 1) and especially worse for longer horizons.

We also considered the impact of our contributions on a version of the Hub ensemble constructed by taking the quantile-wise mean, rather than the median. General trends were similar, with the notable exception of the convolution model, which had a consistently positive impact on the median ensemble, but a mixed and mostly slightly negative impact on the mean ensemble (Fig 4B and S21(B) Fig). This may happen if a model is more correct directionally relative to the pre-existing ensemble, but overshoots in absolute terms, thereby moving the ensemble too far. For both the mean and the median ensemble, changes in performance from adding or removing models were of a similar order of magnitude, suggesting that at least in this instance, with a relatively small ensemble size, the median ensemble was not necessarily more 'robust' to changes than the mean ensemble. However, the ensemble version with all our forecasts included ("hub-ensemble-with-all") tended to perform relatively better for the median ensemble than the mean ensemble, suggesting that adding more models may be more beneficial or 'safer' for the median than for the mean ensemble as directional errors can more easily cancel out than errors in absolute terms.

Discussion

Epidemiological forecasting modelling combines knowledge about infectious disease dynamics with the subjective opinion of the researcher who develops and refines the model. In this study, we compared forecasts of cases of and deaths from COVID-19 in Germany and Poland based purely on human judgement and elicited from a crowd of researchers and volunteers against forecasts from two semi-mechanistic epidemiological models. In spite of the small number of participants and a general tendency to be overconfident, crowd forecasts consistently outperformed our epidemiological models as well as the Hub ensemble when forecasting cases but not when forecasting deaths. This suggests that humans might be relatively good at foreseeing trends that are hard to model but may struggle to form an intuition for the exact relationship between cases and deaths.

Past studies have evaluated the performance of model-based forecasting approaches as well as human experts and non-experts in various contexts. However, most of these studies either focused only on the evaluation of (expert-tuned) model-based approaches [e.g. 12,13,14], or exclusively on human forecasts [19, 20, 24, 25]. In contrast, we directly compared human and

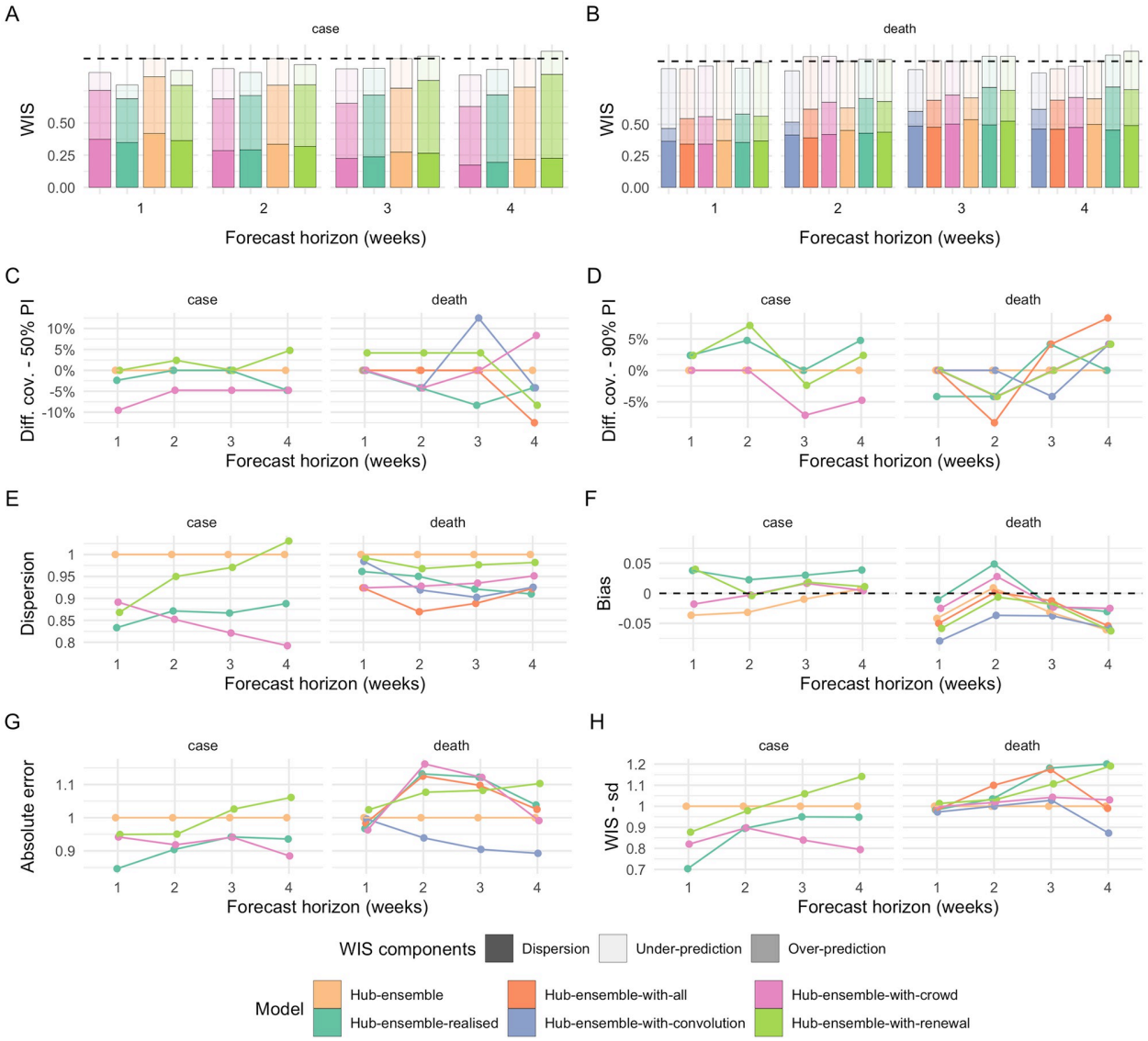

Fig 4. Relative aggregate performance metrics across forecast horizons for different versions of the Hub median ensemble. "Hub-ensemble" *excludes* all our models, Hub-ensemble-all *includes* all of our models, "Hub-ensemble-realised" is the actual hub-ensemble observed in reality, which includes the renewal model and the crowd forecasts, but not the convolution model. A, B: mean weighted interval score (WIS) across horizons relative to the Hub ensemble (lower values indicate better performance). C, D: Empirical coverage of the 50% and 90% prediction intervals minus empirical coverage observed for the Hub ensemble. E: Dispersion relative to the dispersion of the Hub ensemble. Higher values mean greater dispersion of the forecast and imply ceteris paribus a worse score. F: Bias, i.e. general (relative) tendency to over- orunderpredict. Values are between -1 (complete under-prediction) and 1 (complete over-prediction) and 0 ideally. G: Absolute error of the median forecast relative to the Hub ensemble. H. Standard deviation of all WIS values for different horizons relative to the Hub ensemble.

https://doi.org/10.1371/journal.pcbi.1010405.g004

model-based forecasts. This is similar to the approach taken by [11], but extends it in several ways. While Farrow et al. only asked for point predictions and constructed a predictive distribution from these, we asked participants to provide a full predictive distribution, allowing us to compare human forecasts and models without any further assumptions, as well as to analyse how humans quantified their uncertainty. In addition, we compared crowd forecasts to two semi-mechanistic models informed by basic epidemiological knowledge of COVID-19, allowing us to assess not only relative performance but also to analyse qualitative differences

between human judgement and model-based insight. In terms of interpretability of the results, exact knowledge of our two models, as well as focus on a limited set of targets and locations was a major advantage of our study compared to larger studies conducted by the Forecast Hubs [12–15, 17].

The strong performance of crowd forecasts in our study is in line with results from Farrow et al. who also report strong performance of human predictions in past Flu challenges despite difficulties to recruit a large number of participants. The advantage of crowd forecasts we observed over our semi-mechanistic models is likely in part explained by the fact that we compared an ensemble of crowd forecasts with single models. However, this probably explains only part of the difference, and performance relative to the Hub ensemble strongly suggests that human insight is valuable when forecasting highly volatile and potentially hard-to-predict quantities such as case numbers. One potential explanation is that humans can have access to data that is not available to or hard to integrate into model-based forecasts. Relatively good performance of our semi-mechanistic models short-term, but not longer-term, suggests that model-based forecasts are helpful to extrapolate from current conditions, but require some form of human intervention or additional assumptions to inform forecasts when conditions change over time. This human intervention may be particularly important when dealing with artefacts in reporting and data anomalies (and especially when using daily, rather than weekly data). The large variance in predictions in January in Germany for example (severe under-prediction followed by severe over-prediction, see Fig 3A), may in part be caused by the fact that the renewal model operated on daily data and therefore was susceptible to fluctuations in daily reporting which have less of an influence on weekly reporting.

Our results suggest that human intervention may be less beneficial when forecasting deaths (especially at shorter horizons, when deaths are largely dependent on already observed cases), which benefits from the ability to model the delays and exact epidemiological relationships between different leading and lagged indicators. Relatively good performance of the convolution model, especially compared to the poor performance of the renewal model on deaths (which used only deaths to estimate and predict the effective reproduction number) underlines the importance of including leading indicators such as cases as a predictor for deaths.

Given the low number of participants in our study, it is difficult to generalise conclusions about crowd predictions to other settings. Using R shiny as a platform for the web application arguably created some limits to user experience and performance, influencing the number of participants and potentially creating a self-selection effect. Motivating forecasters to contribute regularly proved challenging, especially given that the majority of our participants were from the UK and may not have been familiar with all relevant details of the situation in Germany and Poland. On the other hand, R shiny facilitated quick development and allowed us to provide our crowd forecasting tooling as an open source R package, meaning that it is available for others to use, for example in settings like early-stage outbreaks where model-based forecasts are not available. In light of the relatively small number of Hub ensemble models, performance of the Hub ensemble is also difficult to generalise. More research is needed to replicate these findings and investigate how crowd forecasts compare against the types of models and model ensembles policy makers use to inform their decisions.

Our work suggests that crowd forecasts and model-based forecasts could have different strengths and may be able to complement each other. When choosing a suitable approach for a given task it is important to take into account how the output will be used. In this work we focused on forecasts (which aim to predict future data points whilst accounting for all factors that might influence them), whereas policy makers might be more interested in projections (which show what would happen in the absence of any events that could change the trend) or scenario modelling. Forecasts may not be a suitable basis for informing policy decisions, if

forecasters already have factored in the expectation of a future intervention. Model-based approaches can be either forecasts or projections depending on the assumptions, whereas eliciting projections that are not influenced by implicit assumptions about the future from humans may be harder.

Further work should explore the effects of humans refining their mathematical models or changing model outputs in more detail. Model-based forecasts could be used as an input to human judgement, with researchers adjusting predictions generated by models. Seeing a model-based forecast could help humans calibrate uncertainty better, while allowing for manual intervention to adapt spurious trend predictions. Tools need to be developed to facilitate this process at a larger scale. Human insight could also be used as an input to models. Such a 'hybrid' forecasting approach could for example ask humans to predict the trend of the effective reproduction number R_t or the doubling rate (i.e. how the epidemic evolves) into the future and use this to estimate the exact number of cases, hospitalisations or deaths this would imply. In light of severe overconfidence, yet good performance in terms of the absolute error, post-processing of human forecasts to adjust and widen prediction intervals may be another promising approach. Crowd forecasting in general could benefit greatly from the availability of tools suitable to appeal to a greater audience. Given the good performance we and previous authors observed in spite of the limited resources available and the small number of participants, this seems worthwhile to further develop and explore.

Supporting information

S1 Text. Further details on the semi-mechanistic forecasting models.
(PDF)

S1 Table. Overview of the scoring metrics used.
(PDF)

S2 Table. Scores for one and two week ahead forecasts. Scores are cut to three significant digits and rounded). Note that scores for cases (which include the whole period from October 12th 2020 until March 1st 2021) and deaths (which include only forecasts from the 21st of December 2020 on) are computed on different subsets. Numbers in brackets show the metrics relative to the Hub ensemble (i.e. the median ensemble of all other models submitted to the German and Polish Forecast Hub, excluding our contributions). WIS is the mean weighted interval score (lower values are better), WIS—sd is the standard deviation of all scores achieved by a model. Dispersion, over-prediction and under-prediction together sum up to the weighted interval score. Bias (between -1 and 1, 0 is ideal) represents the general average tendency of a model to over- or underpredict. 50% and 90%-coverage are the percentage of observed values that fell within the 50% and 90% prediction intervals of a model.
(PDF)

S3 Table. Scores for three and four weeks ahead forecasts. Scores are cut to three significant digits and rounded). Note that scores for cases (which include the whole period from October 12th 2020 until March 1st 2021) and deaths (which include only forecasts from the 21st of December 2020 on) are computed on different subsets. Numbers in brackets show the metrics relative to the Hub ensemble (i.e. the median ensemble of all other models submitted to the German and Polish Forecast Hub, excluding our contributions). WIS is the mean weighted interval score (lower values are better), WIS—sd is the standard deviation of all scores achieved by a model. Dispersion, over-prediction and under-prediction together sum up to the weighted interval score. Bias (between -1 and 1, 0 is ideal) represents the general average

tendency of a model to over- or underpredict. 50% and 90%-coverage are the percentage of observed values that fell within the 50% and 90% prediction intervals of a model.
(PDF)

S4 Table. Scores for one and two week ahead forecasts for the different versions of the median ensemble. Scores are cut to three significant digits and rounded. Note that scores for cases (which include the whole period from October 12th 2020 until March 1st 2021) and deaths (which include only forecasts from the 21st of December 2020 on) are computed on different subsets. Numbers in brackets show the metrics relative to the Hub ensemble (i.e. the median ensemble of all other models submitted to the German and Polish Forecast Hub, excluding our contributions). WIS is the mean weighted interval score (lower values are better), WIS—sd is the standard deviation of all scores achieved by a model. Dispersion, over-prediction and under-prediction together sum up to the weighted interval score. Bias (between -1 and 1, 0 is ideal) represents the general average tendency of a model to over- or underpredict. 50% and 90%-coverage are the percentage of observed values that fell within the 50% and 90% prediction intervals of a model.
(PDF)

S5 Table. Scores for three and four week ahead forecasts for the different versions of the median ensemble. Scores are cut to three significant digits and rounded. Note that scores for cases (which include the whole period from October 12th 2020 until March 1st 2021) and deaths (which include only forecasts from the 21st of December 2020 on) are computed on different subsets. Numbers in brackets show the metrics relative to the Hub ensemble (i.e. the median ensemble of all other models submitted to the German and Polish Forecast Hub, excluding our contributions). WIS is the mean weighted interval score (lower values are better), WIS—sd is the standard deviation of all scores achieved by a model. Dispersion, over-prediction and under-prediction together sum up to the weighted interval score. Bias (between -1 and 1, 0 is ideal) represents the general average tendency of a model to over- or underpredict. 50% and 90%-coverage are the percentage of observed values that fell within the 50% and 90% prediction intervals of a model.
(PDF)

S6 Table. Scores for one and two week ahead forecasts for the different versions of the mean ensemble. Scores are cut to three significant digits and rounded. Note that scores for cases (which include the whole period from October 12th 2020 until March 1st 2021) and deaths (which include only forecasts from the 21st of December 2020 on) are computed on different subsets. Numbers in brackets show the metrics relative to the Hub mean ensemble (i.e. the mean ensemble of all other models submitted to the German and Polish Forecast Hub, excluding our contributions). WIS is the mean weighted interval score (lower values are better), WIS—sd is the standard deviation of all scores achieved by a model. Dispersion, over-prediction and under-prediction together sum up to the weighted interval score. Bias (between -1 and 1, 0 is ideal) represents the general average tendency of a model to over- or underpredict. 50% and 90%-coverage are the percentage of observed values that fell within the 50% and 90% prediction intervals of a model.
(PDF)

S7 Table. Scores for three and four week ahead forecasts for the different versions of the mean ensemble. Scores are cut to three significant digits and rounded. Note that scores for cases (which include the whole period from October 12th 2020 until March 1st 2021) and deaths (which include only forecasts from the 21st of December 2020 on) are computed on different subsets. Numbers in brackets show the metrics relative to the Hub mean ensemble (i.e.

the mean ensemble of all other models submitted to the German and Polish Forecast Hub, excluding our contributions). WIS is the mean weighted interval score (lower values are better), WIS—sd is the standard deviation of all scores achieved by a model. Dispersion, over-prediction and under-prediction together sum up to the weighted interval score. Bias (between -1 and 1, 0 is ideal) represents the general average tendency of a model to over- or underpredict. 50% and 90%-coverage are the percentage of observed values that fell within the 50% and 90% prediction intervals of a model.
(PDF)

S8 Table. Scores for one and two week ahead forecasts (sensitivity analysis). Scores are cut to three significant digits and rounded. In the original analysis, cases and deaths were scored on different periods, as the convolution model was only added later. This table shows performance of all models restricted to the period from December 14 2020 until March 1st 2021 where all models were available. Numbers in brackets show the metrics relative to the Hub ensemble (i.e. the median ensemble of all other models submitted to the German and Polish Forecast Hub, excluding our contributions). WIS is the mean weighted interval score (lower values are better), WIS—sd is the standard deviation of all scores achieved by a model. Dispersion, over-prediction and under-prediction together sum up to the weighted interval score. Bias (between -1 and 1, 0 is ideal) represents the general average tendency of a model to over- or underpredict. 50% and 90%-coverage are the percentage of observed values that fell within the 50% and 90% prediction intervals of a model.
(PDF)

S9 Table. Scores for three and four week ahead forecasts (sensitivity analysis). Scores are cut to three significant digits and rounded. In the original analysis, cases and deaths were scored on different periods, as the convolution model was only added later. This table shows performance of all models restricted to the period from December 14 2020 until March 1st 2021 where all models were available. Numbers in brackets show the metrics relative to the Hub ensemble (i.e. the median ensemble of all other models submitted to the German and Polish Forecast Hub, excluding our contributions). WIS is the mean weighted interval score (lower values are better), WIS—sd is the standard deviation of all scores achieved by a model. Dispersion, over-prediction and under-prediction together sum up to the weighted interval score. Bias (between -1 and 1, 0 is ideal) represents the general average tendency of a model to over- or underpredict. 50% and 90%-coverage are the percentage of observed values that fell within the 50% and 90% prediction intervals of a model.
(PDF)

S10 Table. Overview of the models and ensembles used.
(PDF)

S1 Fig. Screenshot of the crowdforecasting app used to elicit predictions (made in June 2021).
(TIF)

S2 Fig. Visualisation of aggregate performance metrics for forecasts one to four weeks into the future in Germany. A, B: mean weighted interval score (WIS, lower indicates better performance) across horizons. WIS is decomposed into its components dispersion, over-prediction and under-prediction. C: Empirical coverage of the 50% prediction intervals (50% coverage is perfect). D: Empirical coverage of the 90% prediction intervals. E: Dispersion (same as in panel A, B). Higher values mean greater dispersion of the forecast and imply ceteris paribus a worse score. F: Bias, i.e. general (relative) tendency to over- or underpredict. Values

are between -1 (complete under-prediction) and 1 (complete over-prediction) and 0 ideally. G: Absolute error of the median forecast (lower is better). H. Standard deviation of all WIS values for different horizons
(TIF)

S3 Fig. Visualisation of aggregate performance metrics for forecasts one to four weeks into the future in Poland. A, B: mean weighted interval score (WIS, lower indicates better performance) across horizons. WIS is decomposed into its components dispersion, over-prediction and under-prediction. C: Empirical coverage of the 50% prediction intervals (50% coverage is perfect). D: Empirical coverage of the 90% prediction intervals. E: Dispersion (same as in panel A, B). Higher values mean greater dispersion of the forecast and imply ceteris paribus a worse score. F: Bias, i.e. general (relative) tendency to over- or underpredict. Values are between -1 (complete under-prediction) and 1 (complete over-prediction) and 0 ideally. G: Absolute error of the median forecast (lower is better). H. Standard deviation of all WIS values for different horizons.
(TIF)

S4 Fig. Visualisation of aggregate performance metrics across locations. A, B: mean weighted interval score (WIS, lower indicates better performance) across horizons. WIS is decomposed into its components dispersion, over-prediction and under-prediction. C: Empirical coverage of the 50% prediction intervals (50% coverage is perfect). D: Empirical coverage of the 90% prediction intervals. E: Dispersion (same as in panel A, B). Higher values mean greater dispersion of the forecast and imply ceteris paribus a worse score. F: Bias, i.e. general (relative) tendency to over- or underpredict. Values are between -1 (complete under-prediction) and 1 (complete over-prediction) and 0 ideally. G: Absolute error of the median forecast (lower is better). H. Standard deviation of WIS values.
(TIF)

S5 Fig. Visualisation of aggregate performance metrics across locations in relative terms. A, B: mean weighted interval score (WIS) across locations (lower values indicate better performance). C, D: Empirical coverage of the 50% and 90% prediction intervals. E: Dispersion. Higher values mean greater dispersion of the forecast and imply ceteris paribus a worse score. F: Bias, i.e. general (relative) tendency to over- orunderpredict. Values are between -1 (complete under-prediction) and 1 (complete over-prediction) and 0 ideally. G: Absolute error of the median forecast. H. Standard deviation of WIS values.
(TIF)

S6 Fig. Visualisation of daily report data. The black line represents weekly data divided by seven. Data were last accessed through the German and Polish Forecast Hub on August 21 2021.
(TIF)

S7 Fig. Visualisation of the absolute difference between the daily report data at the time and the data now. In Germany, there were zero cases and deaths reported on 2020–10-12, and only later 2467 cases and 6 deaths were added. Data were last accessed through the German and Polish Forecast Hub on May 10 2022.
(TIF)

S8 Fig. Visualisation of the relative difference between the weekly report data at the time and the data now. Apart from the data that was retrospectively added on 2020–10-12, data updates did not have a noticeable effect on weekly data (as shown in the forecasting

application). Data were last accessed through the German and Polish Forecast Hub on May 10 2022.
(TIF)

S9 Fig. Visualisation of forecasts and scores for one week ahead forecasts. A, C: Visualisation of 50% prediction intervals of one week ahead forecasts against the reported values. Forecasts that were not scored (because there was no complete set of death forecasts available) are greyed out. B, D: Visualisation of corresponding WIS.
(TIF)

S10 Fig. Visualisation of forecasts and scores for three week ahead forecasts. A, C: Visualisation of 50% prediction intervals of three week ahead forecasts against the reported values. Forecasts that were not scored (because there was no complete set of death forecasts available) are greyed out. B, D: Visualisation of corresponding WIS.
(TIF)

S11 Fig. Visualisation of forecasts and scores for three week ahead forecasts. A, C: Visualisation of 50% prediction intervals of four week ahead forecasts against the reported values. Forecasts that were not scored (because there was no complete set of death forecasts available) are greyed out. B, D: Visualisation of corresponding WIS.
(TIF)

S12 Fig. Distribution of weighted interval scores for one week ahead forecasts. A: Distribution of weighted interval scores for one week ahead forecasts of the different models and forecast targets pooled across locations. B: Distribution of WIS separate by country.
(TIF)

S13 Fig. Distribution of weighted interval scores for three week ahead forecasts. A: Distribution of weighted interval scores for three week ahead forecasts of the different models and forecast targets pooled across locations. B: Distribution of WIS separate by country.
(TIF)

S14 Fig. Distribution of weighted interval scores for four week ahead forecasts. A: Distribution of weighted interval scores for four week ahead forecasts of the different models and forecast targets pooled across locations. B: Distribution of WIS separate by country.
(TIF)

S15 Fig. Distribution of model ranks (in terms of WIS) for one week ahead forecasts. A: Distribution of the ranks (determined by the weighted interval score) for one week ahead forecasts of the different models and forecast targets, pooled across locations. B: Distribution of ranks separate by country.
(TIF)

S16 Fig. Distribution of model ranks (in terms of WIS) for two week ahead forecasts. A: Distribution of the ranks (determined by the weighted interval score) for two week ahead forecasts of the different models and forecast targets, pooled across locations. B: Distribution of ranks separate by country.
(TIF)

S17 Fig. Distribution of model ranks (in terms of WIS) for three week ahead forecasts. A: Distribution of the ranks (determined by the weighted interval score) for three week ahead

forecasts of the different models and forecast targets, pooled across locations. B: Distribution of ranks separate by country.
(TIF)

S18 Fig. Distribution of model ranks (in terms of WIS) for four week ahead forecasts. A: Distribution of the ranks (determined by the weighted interval score) for four week ahead forecasts of the different models and forecast targets, pooled across locations. B: Distribution of ranks separate by country.
(TIF)

S19 Fig. Difference in WIS between the Crowd forecast and the Hub ensemble for two week ahead forecasts. Values below zero mean better performance of the Crowd forecasts.
(TIF)

S20 Fig. Difference in WIS between the Crowd forecast and the Renewal model for two week ahead forecasts. Values below zero mean better performance of the Crowd forecasts.
(TIF)

S21 Fig. Visualisation of aggregate performance metrics across forecast horizons for the different versions of the Hub mean ensemble. "Hub-ensemble" *excludes* all our models, Hub-ensemble-all *includes* all of our models, "Hub-ensemble-realised" is the actual hub-ensemble observed in reality, which includes the renewal model and the crowd forecasts, but ont the convolution model. Values (except for Bias) are computed as differences to the Hub ensemble which excludes our contributions. For Coverage, this is an absolute difference, for other metrics this is a percentage difference. A, B: mean weighted interval score (WIS) across horizons relative to the Hub ensemble (lower values indicate better performance). C, D: Empirical coverage of the 50% and 90% prediction intervals minus empirical coverage observed for the Hub ensemble. E: Dispersion relative to the dispersion of the Hub ensemble. Higher values mean greater dispersion of the forecast and imply ceteris paribus a worse score. F: Bias, i.e. general (relative) tendency to over- orunderpredict. Values are between -1 (complete under-prediction) and 1 (complete over-prediction) and 0 ideally. G: Absolute error of the median forecast relative to the Hub ensemble. H. Standard deviation of all WIS values for different horizons relative to the Hub ensemble.
(TIF)

S22 Fig. Visualisation of aggregate performance metrics across forecast horizons (period from December 14th 2020 on). From December 14th 2020 on, all models were available. In the original analysis, cases and deaths were scored on different periods, as the convolution model was only added later. This sensitivity analysis shows performance of all models restricted to the period from December 14 2020 until March 1st 2021 where all models were available. A, B: mean weighted interval score (WIS, lower indicates better performance) across horizons. WIS is decomposed into its components dispersion, over-prediction and under-prediction. C: Empirical coverage of the 50% prediction intervals (50% coverage is perfect). D: Empirical coverage of the 90% prediction intervals. E: Dispersion (same as in panel A, B). Higher values mean greater dispersion of the forecast and imply ceteris paribus a worse score. F: Bias, i.e. general (relative) tendency to over- or underpredict. Values are between -1 (complete under-prediction) and 1 (complete over-prediction) and 0 ideally. G: Absolute error of the median forecast (lower is better). H. Standard deviation of all WIS values for different horizons
(TIF)

S23 Fig. Number of participants who submitted a forecast over time.
(TIF)

S24 Fig. Number of member models in the official Hub ensemble. This includes our crowd forecasts and the renewal model. Note that the renewal model was not included in the ensemble on December 28th 2020.
(TIF)

S25 Fig. Crowd forecasts and baseline shown in the application for a two week horizon. Shown are the median, as well as the 50% and 90% prediction intervals (in order of decreasing opacity). For any given point in time, the baseline shown in red is what forecasters saw when they opened the app (the baseline shown was constant across all forecast horizons).
(TIF)

S1 Acknowledgements. Members of the CMMID COVID-19 working group.
(PDF)

Acknowledgments

We thank all forecasters who participated in this study for their contribution.

Author Contributions

Conceptualization: Nikos I. Bosse, Sam Abbott, Johannes Bracher, Edwin van Leeuwen, Anne Cori, Sebastian Funk.

Data curation: Nikos I. Bosse.

Formal analysis: Nikos I. Bosse, Sam Abbott.

Investigation: Nikos I. Bosse, Sam Abbott.

Methodology: Nikos I. Bosse, Sam Abbott.

Software: Nikos I. Bosse, Sam Abbott.

Supervision: Sam Abbott, Johannes Bracher, Edwin van Leeuwen, Anne Cori, Sebastian Funk.

Visualization: Nikos I. Bosse.

Writing – original draft: Nikos I. Bosse.

Writing – review & editing: Nikos I. Bosse, Sam Abbott, Johannes Bracher, Habakuk Hain, Billy J. Quilty, Mark Jit, Edwin van Leeuwen, Anne Cori, Sebastian Funk.

References

1. McGowan CJ, Biggerstaff M, Johansson M, Apfeldorf KM, Ben-Nun M, Brooks L, et al. Collaborative efforts to forecast seasonal influenza in the United States, 2015–2016. Scientific Reports. 2019; 9: 683. https://doi.org/10.1038/s41598-018-36361-9 PMID: 30679458
2. Reich NG, Brooks LC, Fox SJ, Kandula S, McGowan CJ, Moore E, et al. A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States. PNAS. 2019; 116: 3146–3154. https://doi.org/10.1073/pnas.1812594116 PMID: 30647115
3. Shaman J, Karspeck A. Forecasting seasonal outbreaks of influenza. PNAS. 2012; 109: 20425–20430. https://doi.org/10.1073/pnas.1208772109 PMID: 23184969
4. Biggerstaff M, Alper D, Dredze M, Fox S, Fung IC-H, Hickmann KS, et al. Results from the centers for disease control and prevention's predict the 2013–2014 Influenza Season Challenge. BMC Infectious Diseases. 2016; 16: 357. https://doi.org/10.1186/s12879-016-1669-x PMID: 27449080

5. Johansson MA, Apfeldorf KM, Dobson S, Devita J, Buczak AL, Baugher B, et al. An open challenge to advance probabilistic forecasting for dengue epidemics. PNAS. 2019; 116: 24268–24274. https://doi.org/10.1073/pnas.1909865116 PMID: 31712420
6. Yamana TK, Kandula S, Shaman J. Superensemble forecasts of dengue outbreaks. Journal of The Royal Society Interface. 2016; 13: 20160410. https://doi.org/10.1098/rsif.2016.0410 PMID: 27733698
7. Colón-González FJ, Bastos LS, Hofmann B, Hopkin A, Harpham Q, Crocker T, et al. Probabilistic seasonal dengue forecasting in Vietnam: A modelling study using superensembles. PLOS Medicine. 2021; 18: e1003542. https://doi.org/10.1371/journal.pmed.1003542
8. Viboud C, Sun K, Gaffey R, Ajelli M, Fumanelli L, Merler S, et al. The RAPIDD ebola forecasting challenge: Synthesis and lessons learnt. Epidemics. 2018; 22: 13–21. https://doi.org/10.1016/j.epidem.2017.08.002 PMID: 28958414
9. Funk S, Camacho A, Kucharski AJ, Lowe R, Eggo RM, Edmunds WJ. 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. 2019; 15: e1006785. https://doi.org/10.1371/journal.pcbi.1006785 PMID: 30742608
10. Del Valle SY, McMahon BH, Asher J, Hatchett R, Lega JC, Brown HE, et al. Summary results of the 2014-2015 DARPA Chikungunya challenge. BMC Infectious Diseases. 2018; 18: 245. https://doi.org/10.1186/s12879-018-3124-7 PMID: 29843621
11. Farrow DC, Brooks LC, Hyun S, Tibshirani RJ, Burke DS, Rosenfeld R. A human judgment approach to epidemiological forecasting. PLOS Computational Biology. 2017; 13: e1005248. https://doi.org/10.1371/journal.pcbi.1005248 PMID: 28282375
12. Funk S, Abbott S, Atkins BD, Baguelin M, Baillie JK, Birrell P, et al. Short-term forecasts to inform the response to the Covid-19 epidemic in the UK. medRxiv. 2020; 2020.11.11.20220962. https://doi.org/10.1101/2020.11.11.20220962
13. Cramer E, Reich NG, Wang SY, Niemi J, Hannan A, House K, et al. COVID-19 Forecast Hub: 4 December 2020 snapshot. Zenodo; 2020.
14. Cramer E, Ray EL, Lopez VK, Bracher J, Brennen A, Rivadeneira AJC, et al. Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the US. medRxiv. 2021; 2021.02.03.21250974.
15. Bracher J, Wolffram D, Deuschel J, Görgen K, Ketterer JL, Ullrich A, et al. A pre-registered short-term forecasting study of COVID-19 in Germany and Poland during the second wave. Nat Commun. 12, 5173 (2021). https://doi.org/10.1038/s41467-021-25207-0
16. Bracher J, Wolffram D, Deuschel J, Görgen K, Ketterer JL, Ullrich A, et al. National and subnational short-term forecasting of COVID-19 in Germany and Poland, early 2021. 2021; 2021.11.05.21265810.
17. European Covid-19 Forecast Hub. European Covid-19 Forecast Hub. 2021 [cited 30 May 2021]. Available: https://covid19forecasthub.eu/
18. Reich NG, McGowan CJ, Yamana TK, Tushar A, Ray EL, Osthus D, et al. Accuracy of real-time multimodel ensemble forecasts for seasonal influenza in the U.S. PLOS Computational Biology. 2019; 15: e1007486. https://doi.org/10.1371/journal.pcbi.1007486 PMID: 31756193
19. Tetlock PE, Mellers BA, Rohrbaugh N, Chen E. Forecasting Tournaments: Tools for Increasing Transparency and Improving the Quality of Debate. Curr Dir Psychol Sci. 2014; 23: 290–295. https://doi.org/10.1177/0963721414534257
20. Atanasov P, Rescober P, Stone E, Swift SA, Servan-Schreiber E, Tetlock P, et al. Distilling the Wisdom of Crowds: Prediction Markets vs. Prediction Polls. Management Science. 2016; 63: 691–706. https://doi.org/10.1287/mnsc.2015.2374
21. Hoogeveen S, Sarafoglou A, Wagenmakers E-J. Laypeople Can Predict Which Social-Science Studies Will Be Replicated Successfully. Advances in Methods and Practices in Psychological Science. 2020; 3: 267–285. https://doi.org/10.1177/2515245920919667
22. ReplicationMarkets. Replication Markets—Reliable research replicates. . .you can bet on it. 2020 [cited 13 Oct 2021]. Available: https://www.replicationmarkets.com/
23. Servan-Schreiber E, Wolfers J, Pennock DM, Galebach B. Prediction Markets: Does Money Matter? Electronic Markets. 2004; 14: 243–251.
24. McAndrew TC, Reich NG. An expert judgment model to predict early stages of the COVID-19 outbreak in the United States. medRxiv. 2020; 2020.09.21.20196725. https://doi.org/10.1101/2020.09.21.20196725 PMID: 32995825
25. Recchia G, Freeman ALJ, Spiegelhalter D. How well did experts and laypeople forecast the size of the COVID-19 pandemic? PLOS ONE. 2021; 16: e0250935. https://doi.org/10.1371/journal.pone.0250935 PMID: 33951092

26. Metaculus. A Preliminary Look at Metaculus and Expert Forecasts. 22 Jun 2020 [cited 30 May 2021]. Available: https://www.metaculus.com/news/2020/06/02/LRT/
27. Hypermind. Hypermind | Supercollective intelligence for decision makers. Hypermind; 2021 [cited 13 Oct 2021]. Available: https://www.hypermind.com/en/
28. CSET Foretell. CSET Foretell. 2021 [cited 13 Oct 2021]. Available: https://www.cset-foretell.com/
29. PredictIt. PredictIt. 2021 [cited 13 Oct 2021]. Available: https://www.predictit.org/
30. Held L, Meyer S, Bracher J. Probabilistic forecasting in infectious disease epidemiology: The 13th Armitage lecture. Statistics in Medicine. 2017; 36: 3443–3460. https://doi.org/10.1002/sim.7363 PMID: 28656694
31. Bosse NI, Abbott S, EpiForecasts, Funk S. Crowdforecastr: Eliciting crowd forecasts in r shiny. 2020. Available: https://github.com/epiforecasts/crowdforecastr.
32. Bosse NI, Abbott S, EpiForecasts, Funk S. Covid.german.forecasts: Forecasting covid-19 related metrics for the german/poland forecast hub. 2020. Available: https://github.com/epiforecasts/covid.german.forecasts
33. ECDC. Download historical data (to 14 December 2020) on the daily number of new reported COVID-19 cases and deaths worldwide. European Centre for Disease Prevention and Control; 14 Dec 2020 [cited 30 May 2021]. Available: https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide
34. RKI. RKI—Coronavirus SARS-CoV-2—Aktueller Lage-/Situationsbericht des RKI zu COVID-19. 2021 [cited 30 May 2021]. Available: https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Situationsberichte/Gesamt.html
35. Forsal.pl. Rozbieżności w statystykach koronawirusa. 22 tys. przypadków będą doliczone do ogólnej liczby wyników. 2020 [cited 30 May 2021]. Available: https://forsal.pl/lifestyle/zdrowie/artykuly/8017628,rozbieznosci-w-statystykach-koronawirusa-22-tys-przypadkow-beda-doliczone-do-ogolnej-liczby-wynikow.html
36. Ärzteblatt DÄG Redaktion Deutsches. SARS-CoV-2-Diagnostik: RKI passt Testempfehlungen an. Deutsches Ärzteblatt; 3 Nov 2020 [cited 30 May 2021]. Available: https://www.aerzteblatt.de/nachrichten/118001/SARS-CoV-2-Diagnostik-RKI-passt-Testempfehlungen-an
37. Fay C, Guyader V, Rochette S, Girard C. Golem: A framework for robust shiny applications. 2021. Available: https://github.com/ThinkR-open/golem
38. Chang W, Cheng J, Allaire J, Sievert C, Schloerke B, Xie Y, et al. Shiny: Web application framework for r. 2021. Available: https://CRAN.R-project.org/package=shiny
39. Our World in Data. COVID-19 Data Explorer. Our World in Data; 2020 [cited 30 May 2021]. Available: https://ourworldindata.org/coronavirus-data-explorer
40. Abbott S, Hellewell J, Hickson J, Munday J, Gostic K, Ellis P, et al. EpiNow2: Estimate real-time case counts and time-varying epidemiological parameters. 2020. Available: https://github.com/epiforecasts/EpiNow2.
41. Fraser C. Estimating Individual and Household Reproduction Numbers in an Emerging Epidemic. PLOS ONE. 2007; 2: e758. https://doi.org/10.1371/journal.pone.0000758 PMID: 17712406
42. epiforecasts.io/covid. Covid-19: Temporal variation in transmission during the COVID-19 outbreak. Covid-19; 2020 [cited 30 May 2021]. Available: https://epiforecasts.io/covid/
43. Sherratt K, Abbott S, Meakin SR, Hellewell J, Munday JD, Bosse N, et al. Exploring surveillance data biases when estimating the reproduction number: with insights into subpopulation transmission of COVID-19 in England *Phil*. Trans. R. Soc. B 37620200283320200283 https://doi.org/10.1098/rstb.2020.0283
44. Abbott S, Hellewell J, Thompson RN, Sherratt K, Gibbs HP, Bosse NI, et al. Estimating the time-varying reproduction number of SARS-CoV-2 using national and subnational case counts. Wellcome Open Res 2020, 5:112. https://doi.org/10.12688/wellcomeopenres.16006.2
45. Xu B, Gutierrez B, Hill S, Scarpino S, Loskill A, Wu J, et al. Epidemiological data from the nCoV-2019 outbreak: Early descriptions from publicly available data. 2020. Available: http://virological.org/t/epidemiological-data-from-the-ncov-2019-outbreak-early-descriptions-from-publicly-available-data/337
46. Stan Development Team. RStan: The r interface to stan. 2020. Available: http://mc-stan.org/
47. Bracher J, Ray EL, Gneiting T, Reich NG. Evaluating epidemic forecasts in an interval format. PLoS Comput Biol. 2021; 17: e1008618. https://doi.org/10.1371/journal.pcbi.1008618 PMID: 33577550
48. Gneiting T, Raftery AE. Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association. 2007; 102: 359–378. https://doi.org/10.1198/016214506000001437

49. Bosse NI, Abbott S, EpiForecasts, Funk S. Scoringutils: Utilities for scoring and assessing predictions. 2020. Available: https://epiforecasts.io/scoringutils/.

50. Deutsche Welle. Coronavirus: Germany to impose one-month partial lockdown | DW | 28.10.2020. 2020 [cited 29 Jun 2021]. Available: https://www.dw.com/en/coronavirus-germany-to-impose-one-month-partial-lockdown/a-55421241

(APPENDIX) Supplementary information

Supplementary information

Scoring metrics used

Table S1. Overview of the scoring metrics used.

Metric	Explanation
WIS (Weighted) interval score	The weighted interval score (smaller values are better) is a proper scoring rule for quantile forecasts. It converges to the continuos ranked probability score (which itself is a generalisation of the absolute error to probabilistic forecasts) for an increasing number of intervals. The score can be decomposed into a dispersion (uncertainty) component and penalties for over- and underprediction. For a single interval, the score is computed as $$IS_\alpha(F,y) = (u-l) + \frac{2}{\alpha}\cdot(l-y)\cdot 1(y \leq l) + \frac{2}{\alpha}\cdot(y-u)\cdot 1(y \geq u),$$ where $1()$ is the indicator function, y is the true value, and l and u are the $\frac{\alpha}{2}$ and $1-\frac{\alpha}{2}$ quantiles of the predictive distribution F, i.e. the lower and upper bound of a single prediction interval. For a set of K prediction intervals and the median m, the score is computed as a weighted sum, $$WIS = \frac{1}{K+0.5}\cdot\left(w_0\cdot|y-m| + \sum_{k=1}^{K} w_k \cdot IS_\alpha(F,y)\right),$$ where w_k is a weight for every interval. Usually, $w_k = \frac{\alpha_k}{2}$ and $w_0 = 0.5$. Its proximity to the absolute error means that when averaging across multiple targets (e.g. different weeks), it will be dominated by targets with higher absolute values.
Interval coverage	Interval coverage is a measure of marginal calibration and indicates the proportion of observed values that fall in a given prediction interval range. Nominal coverage represents the percentage of observed values that should ideally be covered (e.g. we would like a 50 percent prediction interval to cover on average 50 percent of the observations), while empirical coverage is the actual percentage of observations covered by a certain prediction interval.

Table S1. Overview of the scoring metrics used. *(continued)*

Metric	Explanation
Bias	(Relative) bias is a measure of the general tendency of a forecaster to over- or underpredict. Values are between -1 and 1 and 0 ideally. For continuous forecasts, bias is given as $$B(F, y) = 1 - 2 \cdot (F(y)),$$ where F is the CDF of the predictive distribution and y is the observed value. For quantile forecasts, $F(y)$ is replaced by a quantile rank. The appropriate quantile rank is determined by whether the median forecast is below or above the true value. We then take the innermost quantile rank for which the quantile is still larger (under-prediction) or smaller (over-prediction) than the observed value. In contrast to the over- and underprediction penalties of the interval score it is bound between 0 and 1 and represents a general tendency of forecasts to be biased rather than the absolute amount of over- and underprediction. It is therefore a more robust measurement.

The crowdforecasting app

Fig S1. Screenshot of the crowdforecasting app used to elicit predictions (made in June 2021).

Text S1. Further details on the semi-mechanistic forecasting models

Renewal equation model

The model was initialised prior to the first observed data point by assuming constant exponential growth for the mean of assumed delays from infection to case report.

$$I_t = I_0 \exp(rt) \tag{1}$$
$$I_0 \sim \mathcal{LN}(\log I_{obs}, 0.2) \tag{2}$$
$$r \sim \mathcal{LN}(r_{obs}, 0.2) \tag{3}$$

Where I_{obs} and r_{obs} are estimated from the first week of observed data. For the time window of the observed data infections were then modelled by weighting previous infections by the generation time and scaling by the instantaneous reproduction number. These infections were then convolved to cases by date (O_t) and cases by date of report (D_t) using log-normal delay distributions. This model can be defined mathematically as follows,

$$\log R_t = \log R_{t-1} + \mathrm{GP}_t \tag{4}$$
$$I_t = R_t \sum_{\tau=1}^{15} w(\tau|\mu_w, \sigma_w) I_{t-\tau} \tag{5}$$
$$O_t = \sum_{\tau=0}^{15} \xi_O(\tau|\mu_{\xi_O}, \sigma_{\xi_O}) I_{t-\tau} \tag{6}$$
$$D_t = \alpha \sum_{\tau=0}^{15} \xi_D(\tau|\mu_{\xi_D}, \sigma_{\xi_D}) O_{t-\tau} \tag{7}$$
$$C_t \sim \mathrm{NB}\left(\omega_{(t \mod 7)} D_t, \phi\right) \tag{8}$$

Where,

$$w \sim \mathcal{G}(\mu_w, \sigma_w) \tag{9}$$
$$\xi_O \sim \mathcal{LN}(\mu_{\xi_O}, \sigma_{\xi_O}) \tag{10}$$
$$\xi_D \sim \mathcal{LN}(\mu_{\xi_D}, \sigma_{\xi_D}) \tag{11}$$

This model used the following priors for cases,

$$R_0 \sim \mathcal{LN}(0.079, 0.18) \tag{12}$$
$$\mu_w \sim \mathcal{N}(3.6, 0.7) \tag{13}$$
$$\sigma_w \sim \mathcal{N}(3.1, 0.8) \tag{14}$$
$$\mu_{\xi_O} \sim \mathcal{N}(1.62, 0.064) \tag{15}$$
$$\sigma_{\xi_O} \sim \mathcal{N}(0.418, 0.069) \tag{16}$$
$$\mu_{\xi_D} \sim \mathcal{N}(0.614, 0.066) \tag{17}$$
$$\sigma_{\xi_D} \sim \mathcal{N}(1.51, 0.048) \tag{18}$$
$$\alpha \sim \mathcal{N}(0.25, 0.05) \tag{19}$$
$$\frac{\omega}{7} \sim \mathrm{Dirichlet}(1, 1, 1, 1, 1, 1, 1) \tag{20}$$
$$\phi \sim \frac{1}{\sqrt{\mathcal{N}(0, 1)}} \tag{21}$$

and updated the reporting process as follows when forecasting deaths,

$$\mu_{\xi_D} \sim \mathcal{N}(2.29, 0.076) \tag{22}$$
$$\sigma_{\xi_D} \sim \mathcal{N}(0.76, 0.055) \tag{23}$$
$$\alpha \sim \mathcal{N}(0.005, 0.0025) \tag{24}$$

α, μ, σ, and ϕ were truncated to be greater than 0 and with ξ, and w normalised to sum to 1.

The prior for the generation time was sourced from [51] but refit using a log-normal incubation period with a mean of 5.2 days (SD 1.1) and SD of 1.52 days (SD 1.1) with this incubation period also being used as a prior [52] for ξ_O. This resulted in a gamma-distributed generation time with mean 3.6 days (standard deviation (SD) 0.7), and SD of 3.1 days (SD 0.8) for all estimates. We estimated the delay between symptom onset and case report or death required to convolve latent infections to observations by fitting an integer adjusted log-normal distribution to 10 subsampled bootstraps of a public linelist for cases in Germany from April 2020 to June 2020 with each bootstrap using 1% or 1769 samples of the available data [45,53] and combining the posteriors for the mean and standard deviation of the log-normal distribution [40,42,46,54].

GP_t is an approximate Hilbert space Gaussian process as defined in [55] using a Matern 3/2 kernel using a boundary factor of 1.5 and 17 basis functions (20% of the number of days used in fitting). The length scale of the Gaussian process was given a log-normal prior with a mean of 21 days, and a standard deviation of 7 days truncated to be greater than 3 days and less than 60 days. The magnitude of the Gaussian process was assumed to be normally distributed centred at 0 with a standard deviation of 0.1.

From the forecast time horizon (T) and onwards the last value of the Gaussian process was used (hence R_t was assumed to be fixed) and latent infections were adjusted to account for the proportion of the population that was susceptible to infection as follows,

$$I_t = (N - I^c_{t-1}) \left(1 - \exp\left(\frac{-I'_t}{N - I^c_T}\right)\right), \tag{25}$$

where $I^c_t = \sum_{s<t} I_s$ are cumulative infections by $t-1$ and I'_t are the unadjusted infections defined above. This adjustment is based on that implemented in the `epidemia` R package [56,57].

Convolution model The convolution model shares the same observation model as the renewal model but rather than assuming that an observation is predicted by itself using the renewal equation instead assumes that it is predicted entirely by another observation after some parametric delay. It can be defined mathematically as follows,

$$D_t \sim \text{NB}\left(\omega_{(t \mod 7)} \alpha \sum_{\tau=0}^{30} \xi(\tau|\mu, \sigma) C_{t-\tau}, \phi\right) \tag{26}$$

with the following priors,

$$\frac{\omega}{7} \sim \text{Dirichlet}(1, 1, 1, 1, 1, 1, 1) \tag{27}$$

$$\alpha \sim \mathcal{N}(0.01, 0.02) \tag{28}$$

$$\xi \sim \mathcal{LN}(\mu, \sigma) \tag{29}$$

$$\mu \sim \mathcal{N}(2.5, 0.5) \tag{30}$$

$$\sigma \sim \mathcal{N}(0.47, 0.2) \tag{31}$$

$$\phi \sim \frac{1}{\sqrt{\mathcal{N}(0, 1)}} \tag{32}$$

with α, μ, σ, and ϕ truncated to be greater than 0 and with ξ normalised such that $\sum_{\tau=0}^{30} \xi(\tau|\mu, \sigma) = 1$.

Model fitting

Both models were implemented using the `EpiNow2` R package (version 1.3.3) [40]. Each forecast target was fitted independently for each model using Markov-chain Monte Carlo (MCMC) in stan [46]. A minimum of 4 chains were used with a warmup of 250 samples for the renewal equation-based model and 1000 samples for the convolution model. 2000 samples total post warmup were used for the renewal equation model and 4000 samples for the convolution model. Different settings were chosen for each model to optimise compute time contingent on convergence. Convergence was assessed using the R hat diagnostic [46]. For the convolution model forecast the case forecast from the renewal equation model was used in place of observed cases beyond the forecast horizon using 1000 posterior samples. 12 weeks of data was used for both models though only 3 weeks of data were included in the likelihood for the convolution model.

Tables with results of the forecast evaluation

Table S2. Scores for one and two week ahead forecasts (cut to three significant digits and rounded). Note that scores for cases (which include the whole period from October 12th 2020 until March 1st 2021) and deaths (which include only forecasts from the 21st of December 2020 on) are computed on different subsets. Numbers in brackets show the metrics relative to the Hub ensemble (i.e. the median ensemble of all other models submitted to the German and Polish Forecast Hub, excluding our contributions). WIS is the mean weighted interval score (lower values are better), WIS - sd is the standard deviation of all scores achieved by a model. Dispersion, over-prediction and under-prediction together sum up to the weighted interval score. Bias (between -1 and 1, 0 is ideal) represents the general average tendency of a model to over- or underpredict. 50% and 90%-coverage are the percentage of observed values that fell within the 50% and 90% prediction intervals of a model.

	Model	WIS	WIS - sd	dispersion	Underpred.	Overpred.	Bias	Abs. error	50%-Cov.	90%-Cov.
Cases										
1 wk ahead	Crowd forecast	7010 (0.8)	7480 (0.64)	2680 (0.73)	1700 (1.38)	2630 (0.68)	-0.01	10400 (0.82)	0.55	0.79
	Hub-ensemble	8770 (1)	11700 (1)	3670 (1)	1230 (1)	3870 (1)	-0.04	12700 (1)	0.57	0.81
	Renewal	8740 (1)	11800 (1.01)	2190 (0.6)	2720 (2.21)	3830 (0.99)	0.18	12000 (0.94)	0.48	0.71
2 wk ahead	Crowd forecast	16200 (0.89)	16600 (0.76)	3660 (0.6)	5930 (1.56)	6600 (0.78)	-0.01	23300 (0.87)	0.36	0.55
	Hub-ensemble	18300 (1)	21900 (1)	6140 (1)	3800 (1)	8410 (1)	-0.03	26800 (1)	0.43	0.64
	Renewal	25600 (1.4)	33800 (1.54)	5420 (0.88)	5920 (1.56)	14200 (1.69)	0.17	34600 (1.29)	0.43	0.67
Deaths										
1 wk ahead	Convolution	255 (1.03)	343 (1.01)	82 (0.89)	142 (1.23)	31.1 (0.75)	-0.18	399 (1.19)	0.42	0.79
	Crowd forecast	265 (1.07)	317 (0.94)	78.2 (0.85)	82 (0.71)	105 (2.52)	0.08	402 (1.2)	0.38	0.79
	Hub-ensemble	248 (1)	338 (1)	92.2 (1)	115 (1)	41.6 (1)	-0.04	334 (1)	0.62	0.92
	Renewal	298 (1.2)	403 (1.19)	87 (0.94)	107 (0.93)	105 (2.52)	-0.07	413 (1.24)	0.50	0.79
2 wk ahead	Convolution	357 (1.22)	573 (1.49)	104 (0.79)	204 (1.89)	48.8 (0.94)	-0.10	565 (1.32)	0.33	0.79
	Crowd forecast	368 (1.26)	442 (1.15)	107 (0.81)	102 (0.94)	160 (3.08)	0.14	576 (1.34)	0.38	0.75
	Hub-ensemble	292 (1)	385 (1)	132 (1)	108 (1)	51.9 (1)	0.01	429 (1)	0.62	0.96
	Renewal	524 (1.79)	671 (1.74)	155 (1.17)	133 (1.23)	236 (4.55)	-0.02	750 (1.75)	0.50	0.71

Table S3. Scores for three and four week ahead forecasts (cut to three significant digits and rounded). Note that scores for cases (which include the whole period from October 12th 2020 until March 1st 2021) and deaths (which include only forecasts from the 21st of December 2020 on) are computed on different subsets. Numbers in brackets show the metrics relative to the Hub ensemble (i.e. the median ensemble of all other models submitted to the German and Polish Forecast Hub, excluding our contributions). WIS is the mean weighted interval score (lower values are better), WIS - sd is the standard deviation of all scores achieved by a model. Dispersion, over-prediction and under-prediction together sum up to the weighted interval score. Bias (between -1 and 1, 0 is ideal) represents the general average tendency of a model to over- or underpredict. 50% and 90%-coverage are the percentage of observed values that fell within the 50% and 90% prediction intervals of a model.

	Model	WIS	WIS - sd	dispersion	Underpred.	Overpred.	Bias	Abs. error	50%-Cov.	90%-Cov.
Cases										
3 wk ahead	Crowd forecast	27000 (0.81)	26200 (0.64)	4750 (0.52)	11000 (1.43)	11200 (0.67)	0.02	39000 (0.83)	0.14	0.48
	Hub-ensemble	33400 (1)	40700 (1)	9130 (1)	7690 (1)	16600 (1)	-0.01	46900 (1)	0.29	0.62
	Renewal	50600 (1.51)	70000 (1.72)	10800 (1.18)	7710 (1)	32100 (1.93)	0.13	68700 (1.46)	0.29	0.55
4 wk ahead	Crowd forecast	39200 (0.7)	38600 (0.52)	5970 (0.49)	15600 (1.26)	17600 (0.56)	0.07	54800 (0.74)	0.05	0.38
	Hub-ensemble	55900 (1)	73700 (1)	12200 (1)	12400 (1)	31300 (1)	0.01	74400 (1)	0.24	0.52
	Renewal	91700 (1.64)	135000 (1.83)	19500 (1.6)	8990 (0.72)	63200 (2.02)	0.09	125000 (1.68)	0.31	0.48
Deaths										
3 wk ahead	Convolution	541 (1.7)	802 (2.45)	157 (0.91)	279 (3.01)	105 (1.91)	-0.04	747 (1.53)	0.54	0.75
	Crowd forecast	414 (1.3)	526 (1.6)	137 (0.8)	82 (0.88)	194 (3.52)	0.12	648 (1.33)	0.42	0.83
	Hub-ensemble	319 (1)	328 (1)	172 (1)	92.7 (1)	55.1 (1)	-0.03	488 (1)	0.54	0.96
	Renewal	724 (2.27)	916 (2.79)	249 (1.45)	158 (1.7)	317 (5.75)	-0.01	1040 (2.13)	0.46	0.83
4 wk ahead	Convolution	763 (1.8)	932 (2.1)	268 (1.26)	331 (2.63)	164 (1.91)	0.01	985 (1.46)	0.54	0.75
	Crowd forecast	498 (1.17)	633 (1.43)	168 (0.79)	83.6 (0.66)	246 (2.87)	0.14	756 (1.12)	0.38	0.79
	Hub-ensemble	424 (1)	443 (1)	212 (1)	126 (1)	85.7 (1)	-0.06	675 (1)	0.58	0.92
	Renewal	959 (2.26)	1210 (2.73)	337 (1.59)	200 (1.59)	421 (4.91)	-0.05	1350 (2)	0.50	0.79

Aggregate performance by location

Performance in Germany

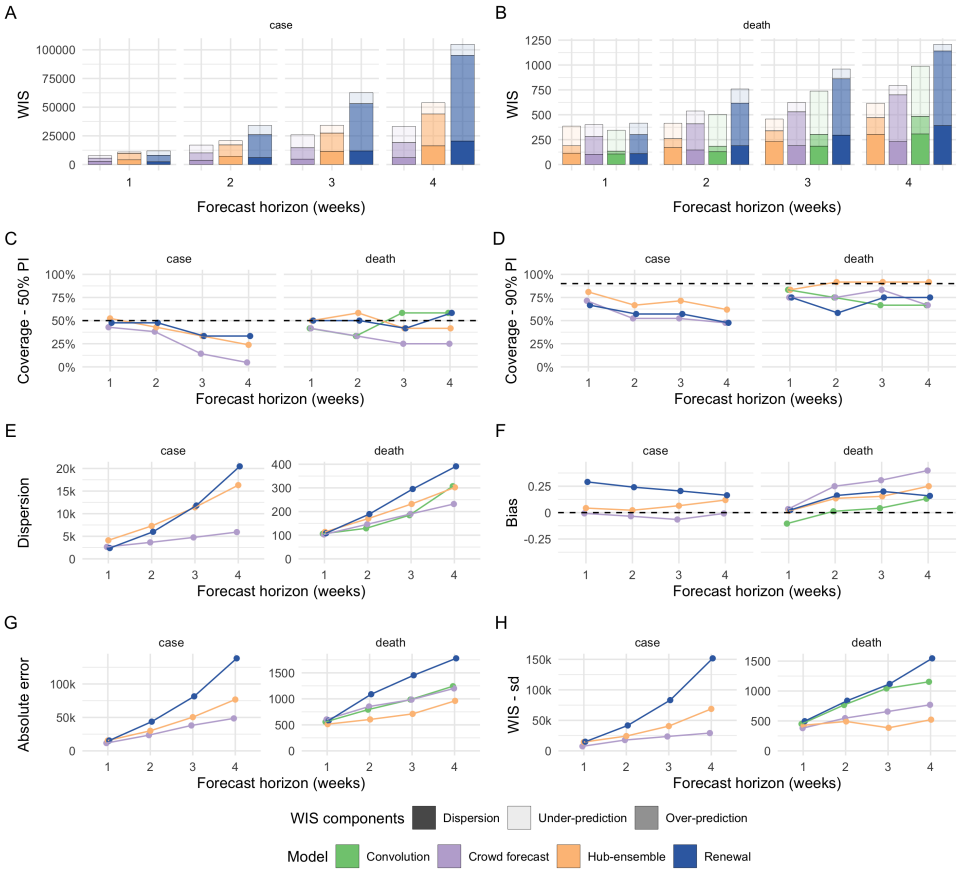

Fig S2. Visualisation of aggregate performance metrics for forecasts one to four weeks into the future in Germany. A, B: mean weighted interval score (WIS, lower indicates better performance) across horizons. WIS is decomposed into its components dispersion, over-prediction and under-prediction. C: Empirical coverage of the 50% prediction intervals (50% coverage is perfect). D: Empirical coverage of the 90% prediction intervals. E: Dispersion (same as in panel A, B). Higher values mean greater dispersion of the forecast and imply ceteris paribus a worse score. F: Bias, i.e. general (relative) tendency to over- or underpredict. Values are between -1 (complete under-prediction) and 1 (complete over-prediction) and 0 ideally. G: Absolute error of the median forecast (lower is better). H. Standard deviation of all WIS values for different horizons

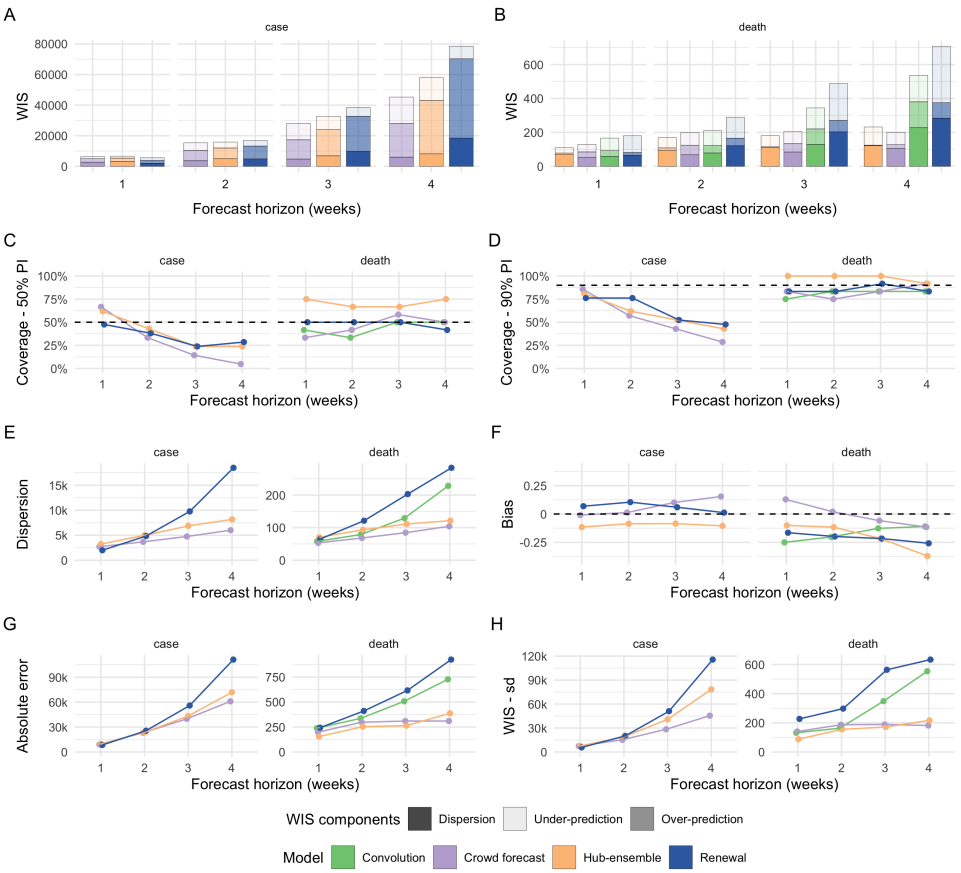

Fig S3. Visualisation of aggregate performance metrics for forecasts one to four weeks into the future in Poland. A, B: mean weighted interval score (WIS, lower indicates better performance) across horizons. WIS is decomposed into its components dispersion, over-prediction and under-prediction. C: Empirical coverage of the 50% prediction intervals (50% coverage is perfect). D: Empirical coverage of the 90% prediction intervals. E: Dispersion (same as in panel A, B). Higher values mean greater dispersion of the forecast and imply ceteris paribus a worse score. F: Bias, i.e. general (relative) tendency to over- or underpredict. Values are between -1 (complete under-prediction) and 1 (complete over-prediction) and 0 ideally. G: Absolute error of the median forecast (lower is better). H. Standard deviation of all WIS values for different horizons

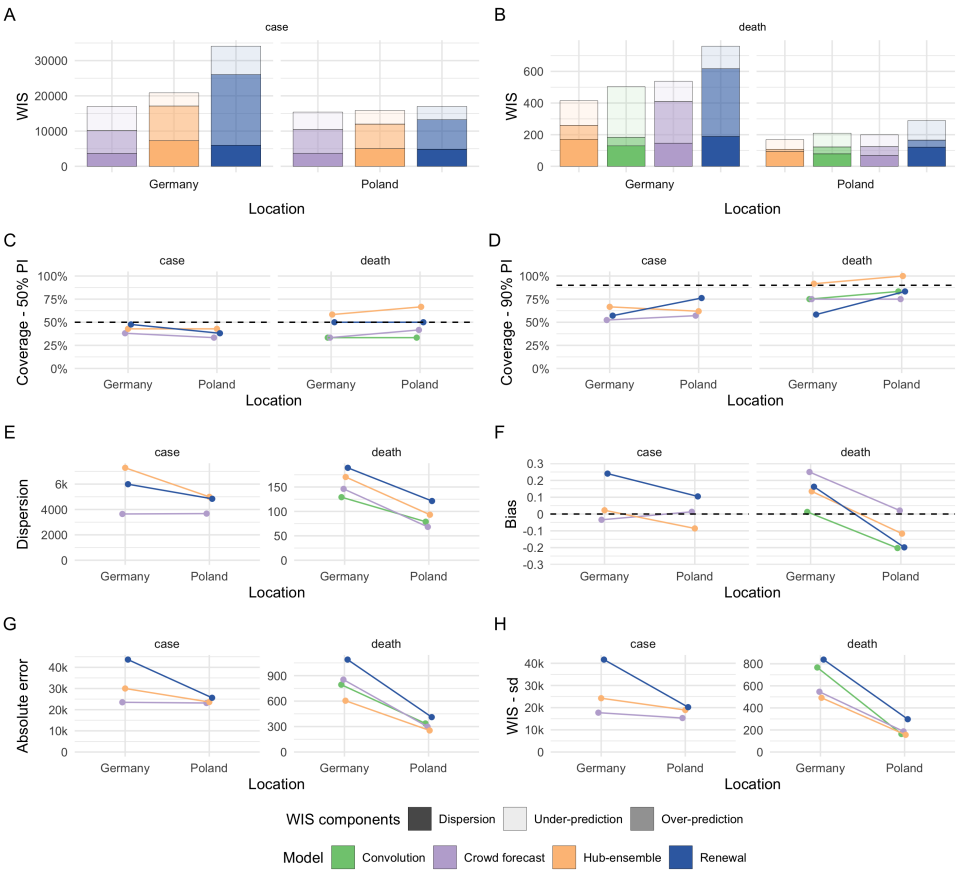

Fig S4. Visualisation of aggregate performance metrics across locations. A, B: mean weighted interval score (WIS, lower indicates better performance) across horizons. WIS is decomposed into its components dispersion, over-prediction and under-prediction. C: Empirical coverage of the 50% prediction intervals (50% coverage is perfect). D: Empirical coverage of the 90% prediction intervals. E: Dispersion (same as in panel A, B). Higher values mean greater dispersion of the forecast and imply ceteris paribus a worse score. F: Bias, i.e. general (relative) tendency to over- or underpredict. Values are between -1 (complete under-prediction) and 1 (complete over-prediction) and 0 ideally. G: Absolute error of the median forecast (lower is better). H. Standard deviation of WIS values.

Performance across locations in relative terms

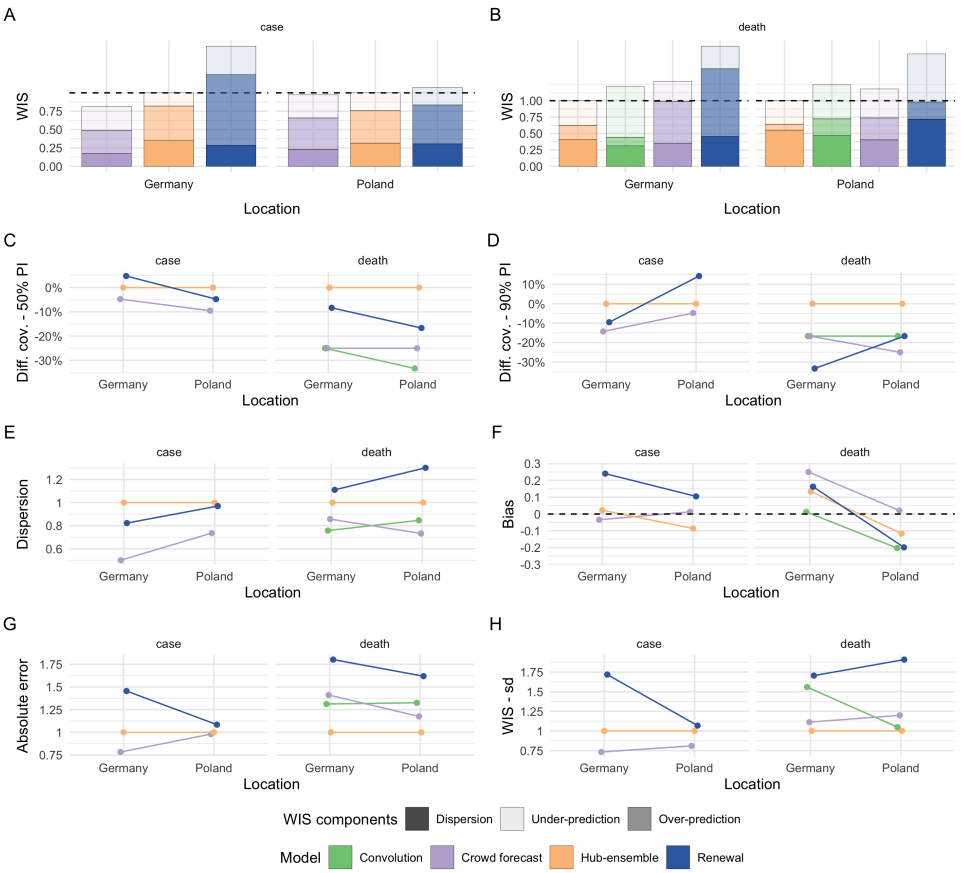

Fig S5. Visualisation of relative aggregate performance metrics across locations. A, B: mean weighted interval score (WIS) across locations (lower values indicate better performance). C, D: Empirical coverage of the 50% and 90% prediction intervals. E: Dispersion. Higher values mean greater dispersion of the forecast and imply ceteris paribus a worse score. F: Bias, i.e. general (relative) tendency to over- orunderpredict. Values are between -1 (complete under-prediction) and 1 (complete over-prediction) and 0 ideally. G: Absolute error of the median forecast. H. Standard deviation of WIS values.

Visualisation of daily reported cases and deaths

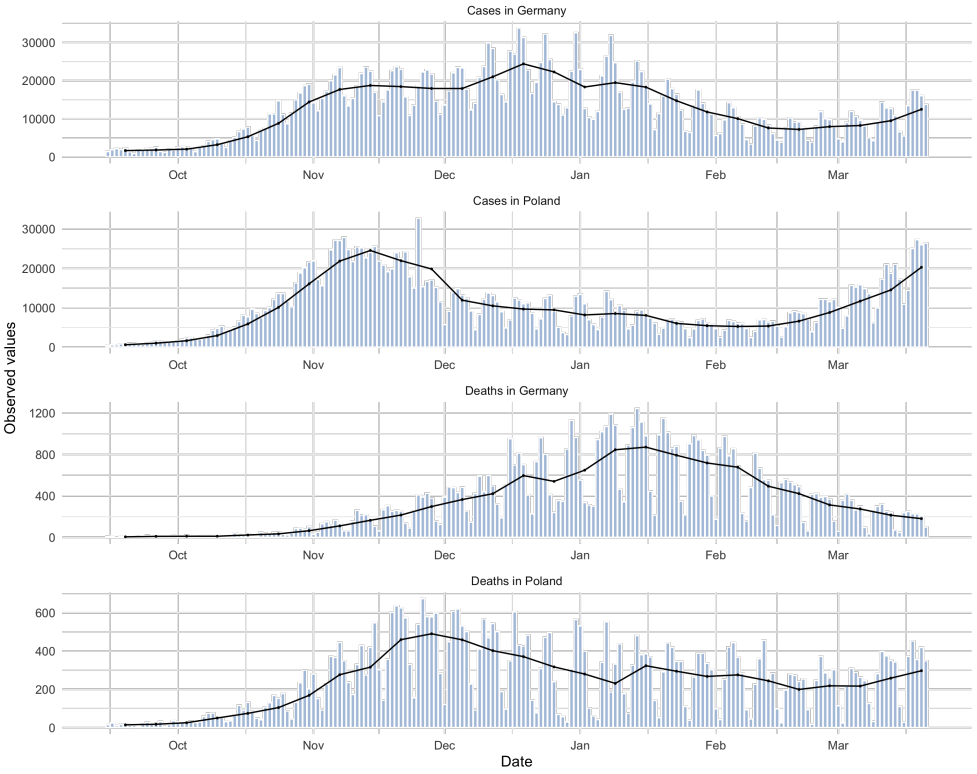

Fig S6. Visualisation of daily report data. The black line represents weekly data divided by seven. Data were last accessed through the German and Polish Forecast Hub on August 21 2021.

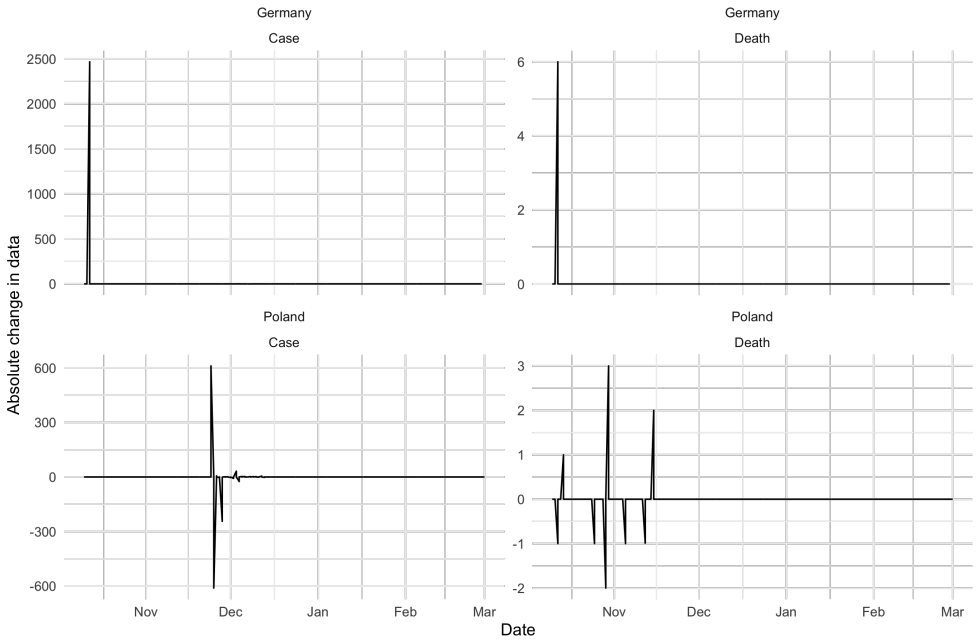

Fig S7. Visualisation of the absolute difference between the daily report data at the time and the data now. In Germany, there were zero cases and deaths reported on 2020-10-12, and only later 2467 cases and 6 deaths were added. Data were last accessed through the German and Polish Forecast Hub on May 10 2022.

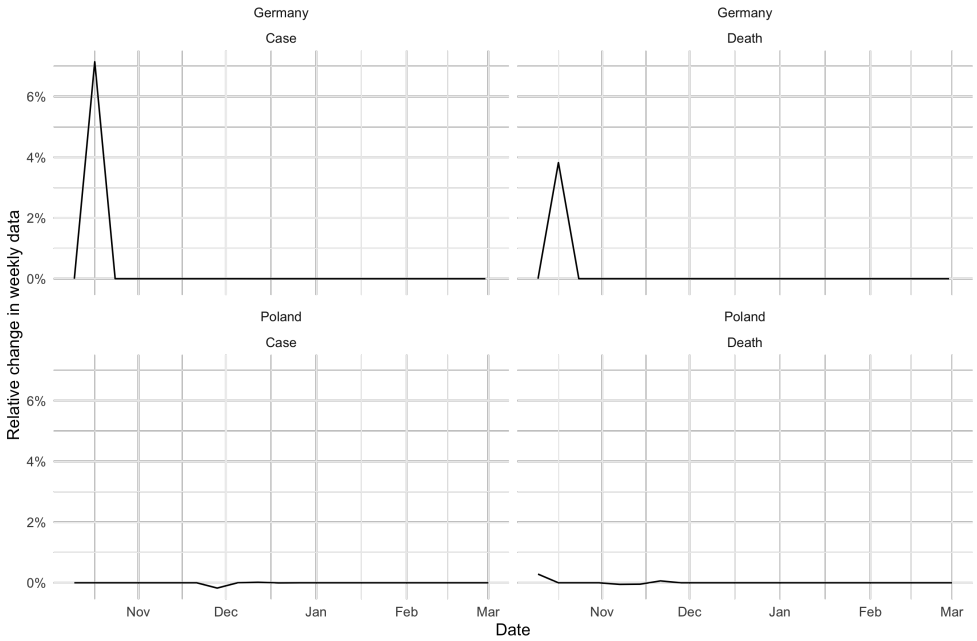

Fig S8. Visualisation of the relative difference between the weekly report data at the time and the data now. Apart from the data that was retrospectively added on 2020-10-12, data updates did not have a noticeable effect on weekly data (as shown in the forecasting application). Data were last accessed through the German and Polish Forecast Hub on May 10 2022.

Fig S9. A, C: Visualisation of 50% prediction intervals of one week ahead forecasts against the reported values. Forecasts that were not scored (because there was no complete set of death forecasts available) are greyed out. B, D: Visualisation of corresponding WIS.

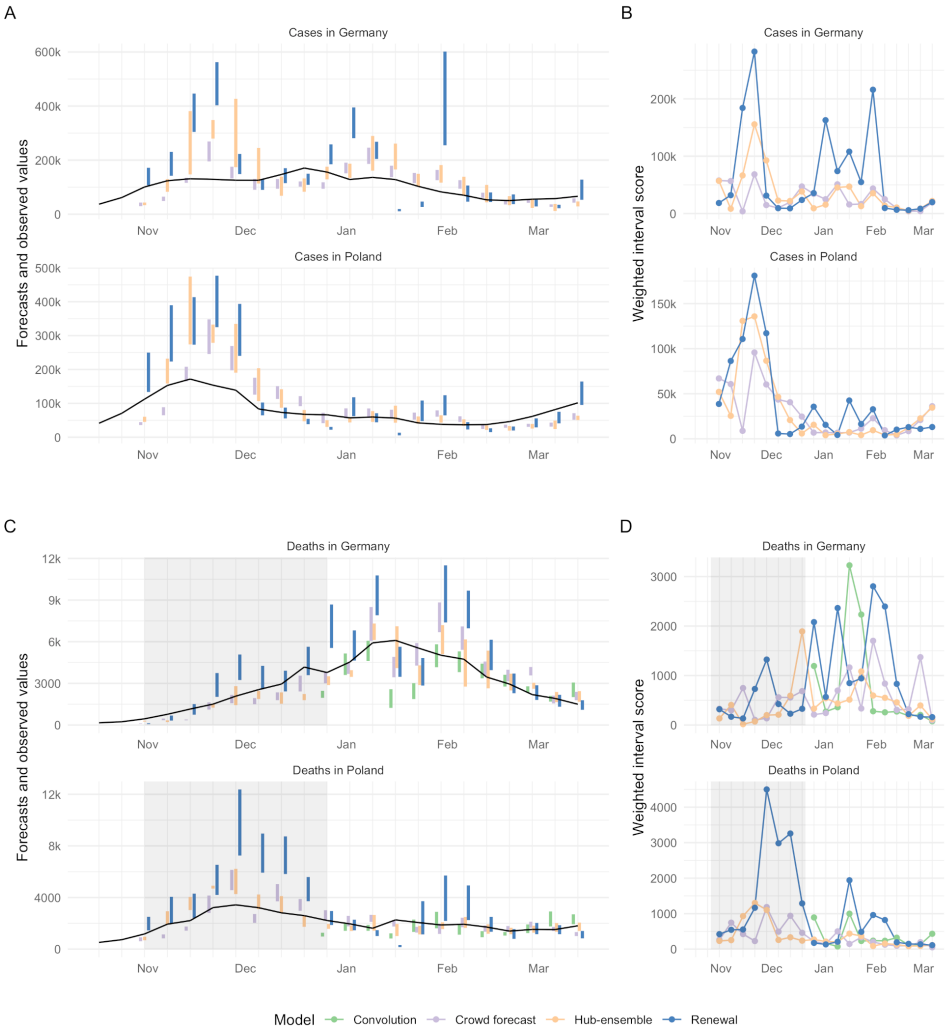

Fig S10. A, C: Visualisation of 50% prediction intervals of three week ahead forecasts against the reported values. Forecasts that were not scored (because there was no complete set of death forecasts available) are greyed out. B, D: Visualisation of corresponding WIS.

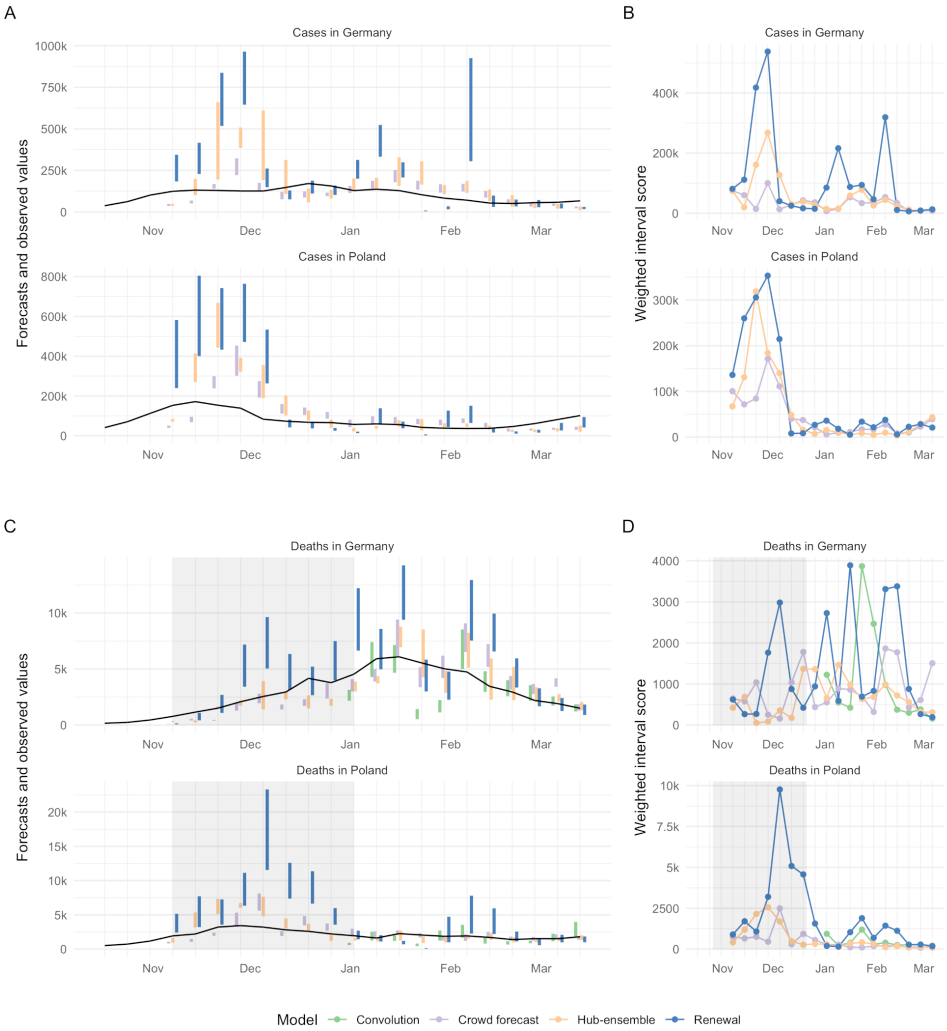

Fig S11. A, C: Visualisation of 50% prediction intervals of four week ahead forecasts against the reported values. Forecasts that were not scored (because there was no complete set of death forecasts available) are greyed out. B, D: Visualisation of corresponding WIS.

Distribution of scores

Absolute scores

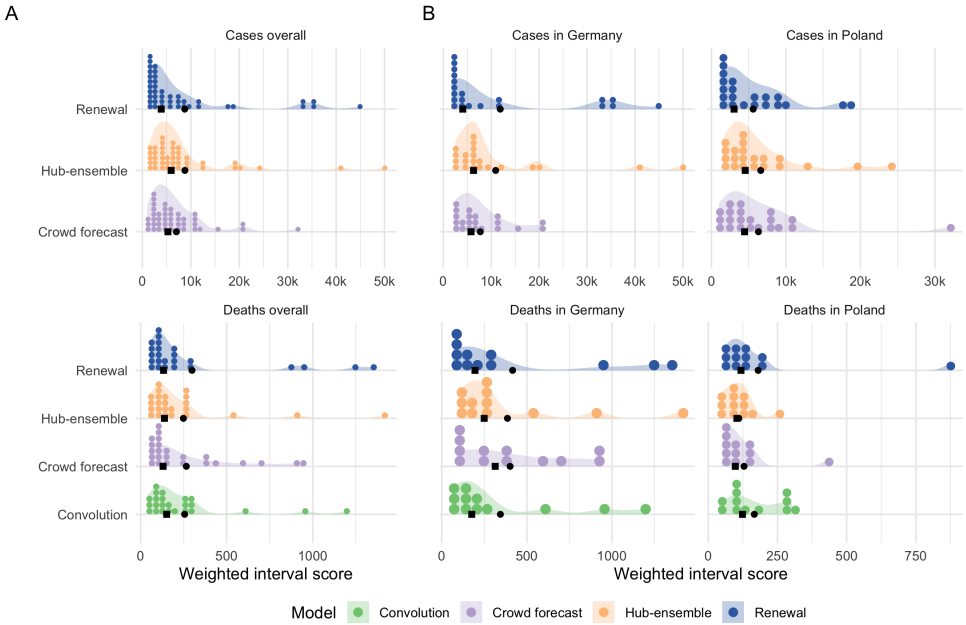

Fig S12. A: Distribution of weighted interval scores for one week ahead forecasts of the different models and forecast targets. B: Distribution of WIS separate by country.

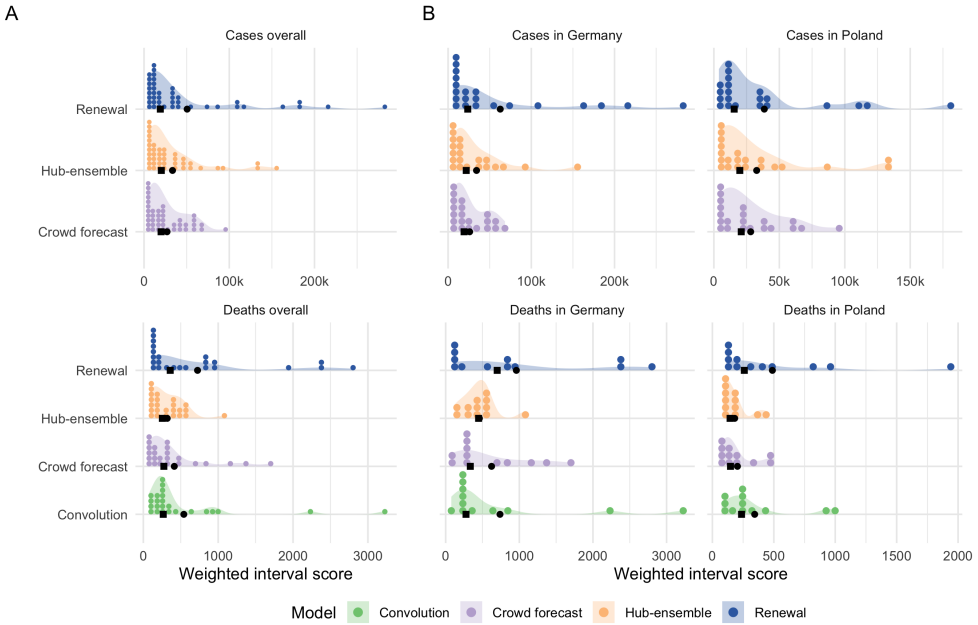

Fig S13. A: Distribution of weighted interval scores for three week ahead forecasts of the different models and forecast targets. B: Distribution of WIS separate by country.

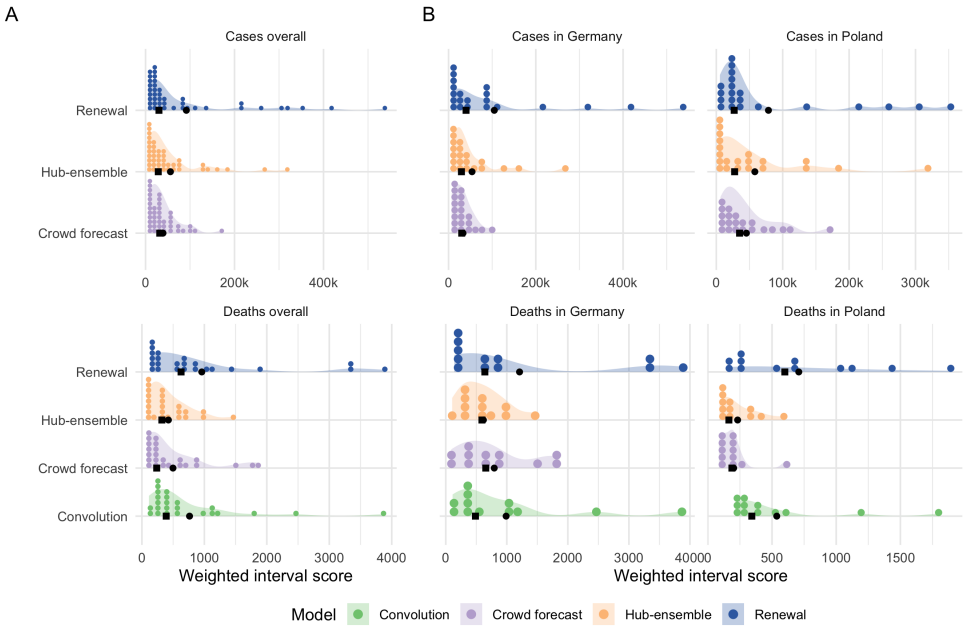

Fig S14. A: Distribution of weighted interval scores for four week ahead forecasts of the different models and forecast targets. B: Distribution of WIS separate by country.

Ranks achieved by forecasts

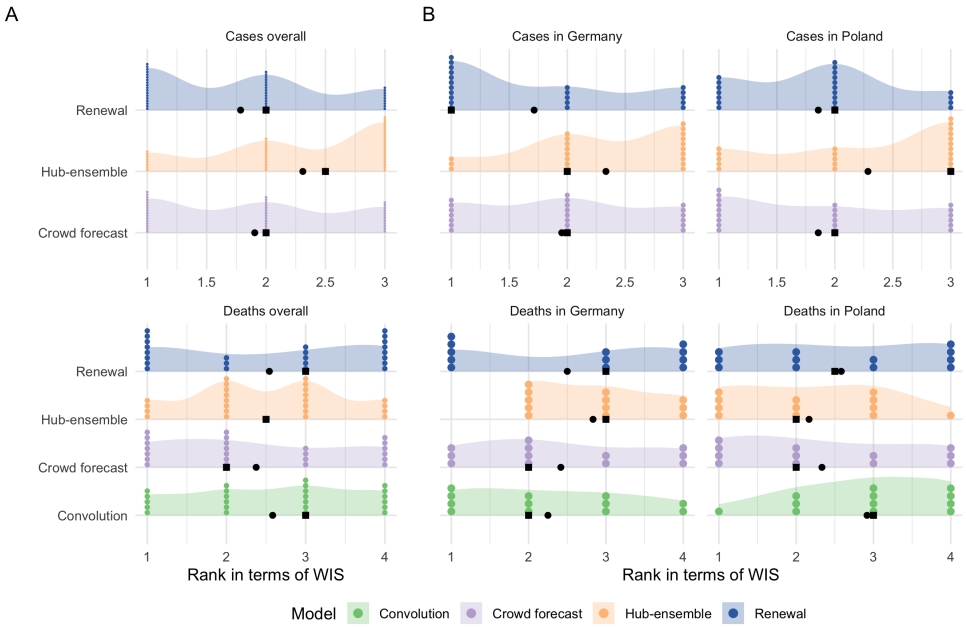

Fig S15. A: Distribution of the ranks (determined by the weighted interval score) for one week ahead forecasts of the different models and forecast targets. B: Distribution of ranks separate by country.

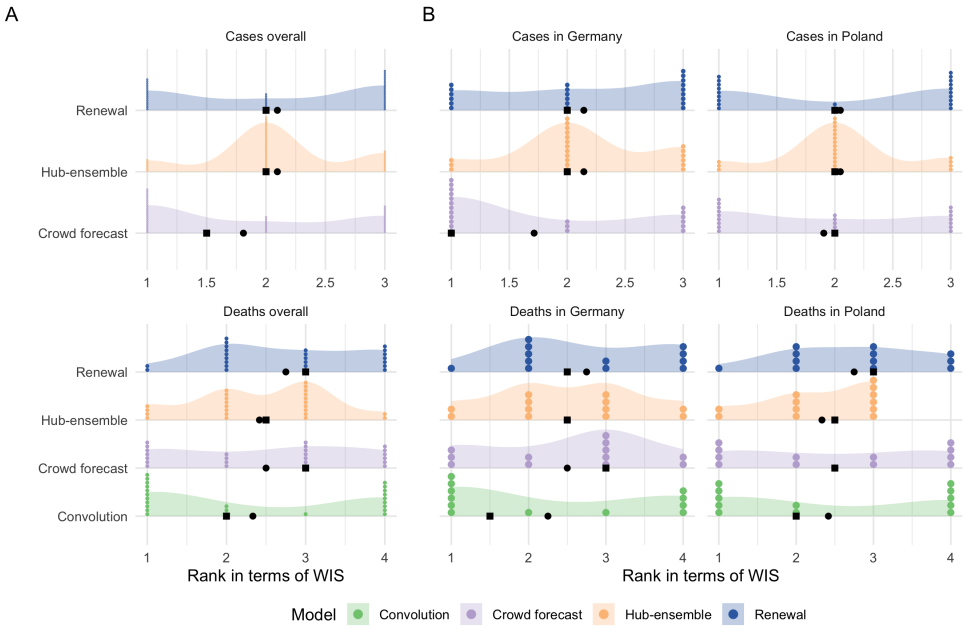

Fig S16. A: Distribution of the ranks (determined by the weighted interval score) for two week ahead forecasts of the different models and forecast targets. B: Distribution of ranks separate by country.

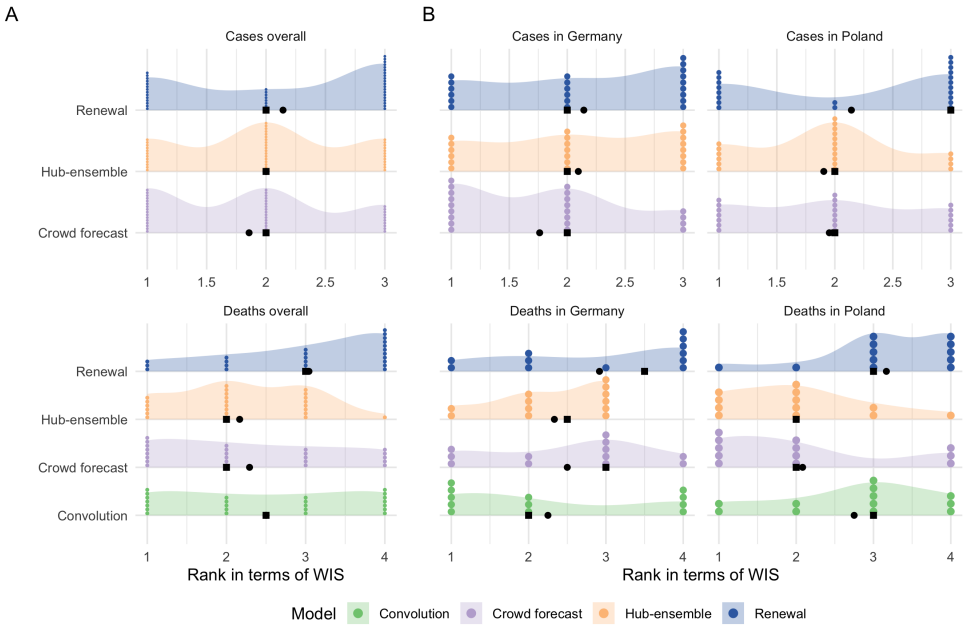

Fig S17. A: Distribution of the ranks (determined by the weighted interval score) for three week ahead forecasts of the different models and forecast targets. B: Distribution of ranks separate by country.

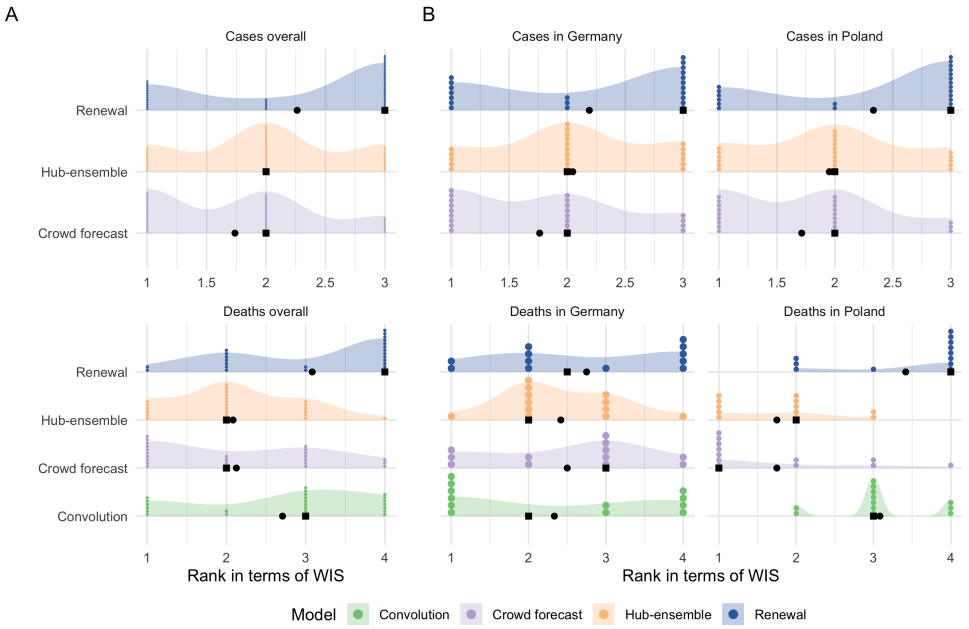

Fig S18. A: Distribution of the ranks (determined by the weighted interval score) for four week ahead forecasts of the different models and forecast targets. B: Distribution of ranks separate by country.

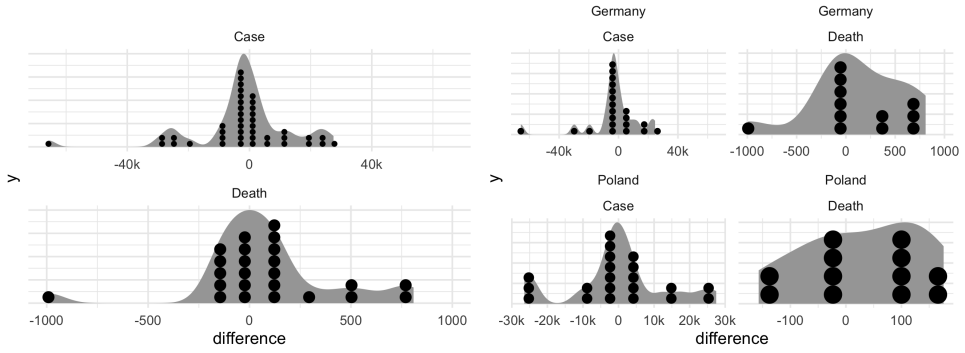

Fig S19. Density plot with the difference in WIS between the Crowd forecast and the Hub ensemble (values below zero mean better performance of the Crowd forecasts) for a 2 week ahead forecast horizon.

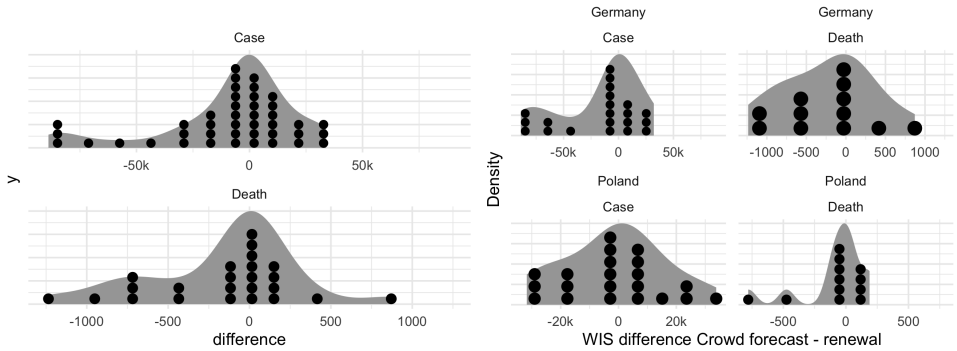

Fig S20. Density plot with the difference in WIS between the Crowd forecast and the Renewal model (values below zero mean better performance of the Crowd forecasts) for a 2 week ahead forecast horizon.

Comparison of ensembles

Performance visualisation mean ensemble

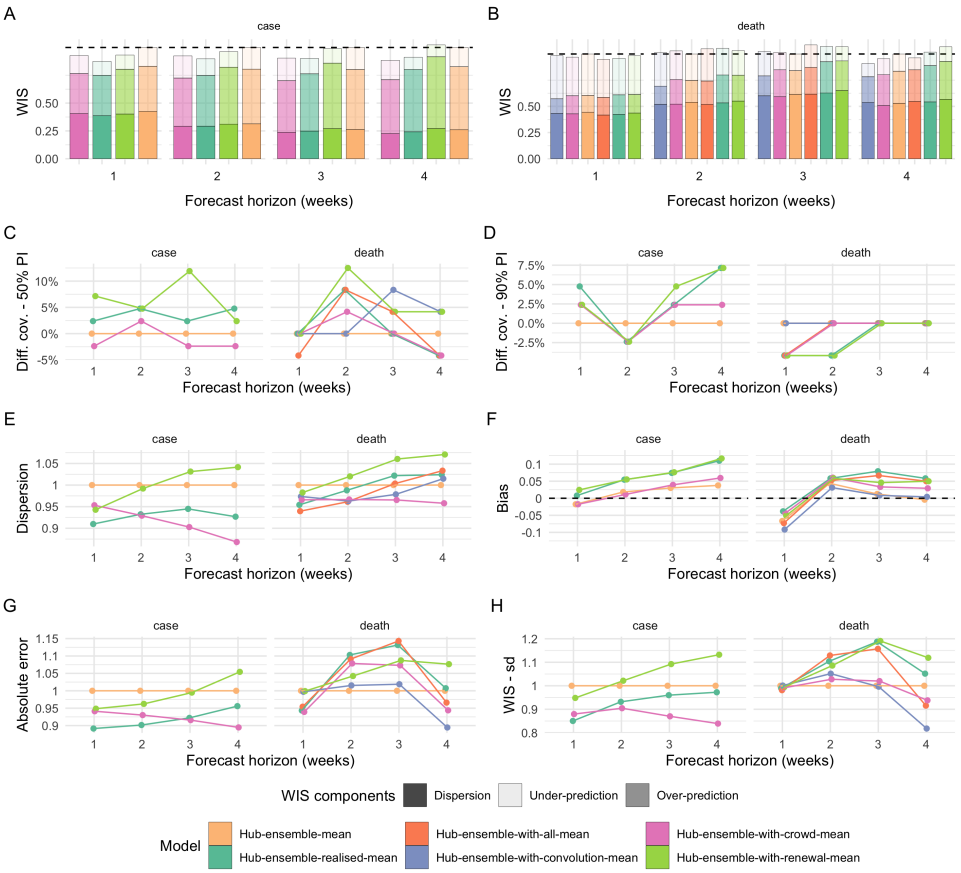

Fig S21. Visualisation of aggregate performance metrics across forecast horizons for the different versions of the Hub mean ensemble. "Hub-ensemble" extitexcludes all our models, Hub-ensemble-all extitincludes all of our models, "Hub-ensemble-realised" is the actual hub-ensemble observed in reality, which includes the renewal model and the crowd forecasts, but ont the convolution model. Values (except for Bias) are computed as differences to the Hub ensemble which excludes our contributions. For Coverage, this is an absolute difference, for other metrics this is a percentage difference. A, B: mean weighted interval score (WIS) across horizons relative to the Hub ensemble (lower values indicate better performance). C, D: Empirical coverage of the 50% and 90% prediction intervals minus empirical coverage observed for the Hub ensemble. E: Dispersion relative to the dispersion of the Hub ensemble. Higher values mean greater dispersion of the forecast and imply ceteris paribus a worse score. F: Bias, i.e. general (relative) tendency to over- orunderpredict. Values are between -1 (complete under-prediction) and 1 (complete over-prediction) and 0 ideally. G: Absolute error of the median forecast relative to the Hub ensemble. H. Standard deviation of all WIS values for different horizons relative to the Hub ensemble.

Tables median ensemble

Tables mean ensemble

Table S4. Scores for one and two week ahead forecasts (cut to three significant digits and rounded) for the different versions of the median ensemble. Note that scores for cases (which include the whole period from October 12th 2020 until March 1st 2021) and deaths (which include only forecasts from the 21st of December 2020 on) are computed on different subsets. Numbers in brackets show the metrics relative to the Hub ensemble (i.e. the median ensemble of all other models submitted to the German and Polish Forecast Hub, excluding our contributions). WIS is the mean weighted interval score (lower values are better), WIS - sd is the standard deviation of all scores achieved by a model. Dispersion, over-prediction and under-prediction together sum up to the weighted interval score. Bias (between -1 and 1, 0 is ideal) represents the general average tendency of a model to over- or underpredict. 50% and 90%-coverage are the percentage of observed values that fell within the 50% and 90% prediction intervals of a model.

	Model	WIS	WIS - sd	dispersion	Underpred.	Overpred.	Bias	Abs. error	50%-Cov.	90%-Cov.
Cases										
1 wk ahead	Hub-ensemble	8770 (1)	11700 (1)	3670 (1)	1230 (1)	3870 (1)	-0.04	12700 (1)	0.57	0.81
	Hub-ensemble-realised	6970 (0.79)	8260 (0.71)	3060 (0.83)	943 (0.77)	2970 (0.77)	0.04	10800 (0.85)	0.55	0.83
	Hub-ensemble-with-crowd	7820 (0.89)	9630 (0.82)	3270 (0.89)	1210 (0.98)	3330 (0.86)	-0.02	12000 (0.94)	0.48	0.81
	Hub-ensemble-with-renewal	7960 (0.91)	10300 (0.88)	3190 (0.87)	1020 (0.83)	3760 (0.97)	0.04	12100 (0.95)	0.57	0.83
2 wk ahead	Hub-ensemble	18300 (1)	21900 (1)	6140 (1)	3800 (1)	8410 (1)	-0.03	26800 (1)	0.43	0.64
	Hub-ensemble-realised	16400 (0.9)	19600 (0.89)	5350 (0.87)	3290 (0.87)	7730 (0.92)	0.02	24200 (0.9)	0.43	0.69
	Hub-ensemble-with-crowd	16900 (0.92)	19600 (0.89)	5230 (0.85)	4310 (1.13)	7370 (0.88)	0.00	24600 (0.92)	0.38	0.64
	Hub-ensemble-with-renewal	17500 (0.96)	21400 (0.98)	5830 (0.95)	2880 (0.76)	8770 (1.04)	0.00	25500 (0.95)	0.45	0.71
Deaths										
1 wk ahead	Hub-ensemble	248 (1)	338 (1)	92.2 (1)	115 (1)	41.6 (1)	-0.04	334 (1)	0.62	0.92
	Hub-ensemble-realised	235 (0.95)	332 (0.98)	88.6 (0.96)	90.4 (0.79)	55.5 (1.33)	-0.01	323 (0.97)	0.62	0.88
	Hub-ensemble-with-all	234 (0.94)	331 (0.98)	85.2 (0.92)	98.1 (0.85)	50.2 (1.21)	-0.05	329 (0.99)	0.62	0.92
	Hub-ensemble-with-convolution	234 (0.94)	329 (0.97)	90.7 (0.98)	118 (1.03)	25.3 (0.61)	-0.08	333 (1)	0.62	0.92
	Hub-ensemble-with-crowd	239 (0.96)	337 (1)	85.2 (0.92)	99.6 (0.87)	54.2 (1.3)	-0.03	322 (0.96)	0.62	0.92
	Hub-ensemble-with-renewal	246 (0.99)	342 (1.01)	91.5 (0.99)	106 (0.92)	48.6 (1.17)	-0.06	342 (1.02)	0.67	0.92
2 wk ahead	Hub-ensemble	292 (1)	385 (1)	132 (1)	108 (1)	51.9 (1)	0.01	429 (1)	0.62	0.96
	Hub-ensemble-realised	296 (1.01)	398 (1.03)	125 (0.95)	91 (0.84)	80.2 (1.55)	0.05	486 (1.13)	0.58	0.92
	Hub-ensemble-with-all	303 (1.04)	423 (1.1)	115 (0.87)	122 (1.13)	66.1 (1.27)	0.00	483 (1.13)	0.62	0.88
	Hub-ensemble-with-convolution	270 (0.92)	385 (1)	121 (0.92)	119 (1.1)	29.9 (0.58)	-0.04	403 (0.94)	0.58	0.96
	Hub-ensemble-with-crowd	303 (1.04)	392 (1.02)	122 (0.92)	106 (0.98)	74.6 (1.44)	0.03	499 (1.16)	0.58	0.92
	Hub-ensemble-with-renewal	296 (1.01)	397 (1.03)	128 (0.97)	97.1 (0.9)	71.2 (1.37)	-0.01	462 (1.08)	0.67	0.92

Table S5. Scores for three and four week ahead forecasts (cut to three significant digits and rounded) for the different versions of the median ensemble. Note that scores for cases (which include the whole period from October 12th 2020 until March 1st 2021) and deaths (which include only forecasts from the 21st of December 2020 on) are computed on different subsets. Numbers in brackets show the metrics relative to the Hub ensemble (i.e. the median ensemble of all other models submitted to the German and Polish Forecast Hub, excluding our contributions). WIS is the mean weighted interval score (lower values are better), WIS - sd is the standard deviation of all scores achieved by a model. Dispersion, over-prediction and under-prediction together sum up to the weighted interval score. Bias (between -1 and 1, 0 is ideal) represents the general average tendency of a model to over- or underpredict. 50% and 90%-coverage are the percentage of observed values that fell within the 50% and 90% prediction intervals of a model.

	Model	WIS	WIS - sd	dispersion	Underpred.	Overpred.	Bias	Abs. error	50%-Cov.	90%-Cov.
Cases										
3 wk ahead	Hub-ensemble	33400 (1)	40700 (1)	9130 (1)	7690 (1)	16600 (1)	-0.01	46900 (1)	0.29	0.62
	Hub-ensemble-realised	30800 (0.92)	38600 (0.95)	7910 (0.87)	6890 (0.9)	16000 (0.96)	0.03	44200 (0.94)	0.29	0.62
	Hub-ensemble-with-crowd	30800 (0.92)	34100 (0.84)	7500 (0.82)	8960 (1.17)	14300 (0.86)	0.02	44100 (0.94)	0.24	0.55
	Hub-ensemble-with-renewal	34000 (1.02)	43100 (1.06)	8860 (0.97)	6300 (0.82)	18900 (1.14)	0.02	48100 (1.03)	0.29	0.60
4 wk ahead	Hub-ensemble	55900 (1)	73700 (1)	12200 (1)	12400 (1)	31300 (1)	0.01	74400 (1)	0.24	0.52
	Hub-ensemble-realised	51200 (0.92)	69900 (0.95)	10900 (0.89)	11100 (0.9)	29300 (0.94)	0.04	69600 (0.94)	0.19	0.57
	Hub-ensemble-with-crowd	48800 (0.87)	58600 (0.8)	9700 (0.8)	13700 (1.1)	25400 (0.81)	0.00	65800 (0.88)	0.19	0.48
	Hub-ensemble-with-renewal	59100 (1.06)	84100 (1.14)	12600 (1.03)	10100 (0.81)	36400 (1.16)	0.01	78900 (1.06)	0.29	0.55
Deaths										
3 wk ahead	Hub-ensemble	319 (1)	328 (1)	172 (1)	92.7 (1)	55.1 (1)	-0.03	488 (1)	0.54	0.96
	Hub-ensemble-realised	332 (1.04)	388 (1.18)	158 (0.92)	78.7 (0.85)	95 (1.72)	-0.02	547 (1.12)	0.46	1.00
	Hub-ensemble-with-all	321 (1.01)	385 (1.17)	153 (0.89)	100 (1.08)	68.1 (1.24)	-0.01	535 (1.1)	0.54	1.00
	Hub-ensemble-with-convolution	298 (0.93)	337 (1.03)	155 (0.9)	106 (1.14)	37.5 (0.68)	-0.04	441 (0.9)	0.67	0.92
	Hub-ensemble-with-crowd	319 (1)	342 (1.04)	160 (0.93)	85.1 (0.92)	73.6 (1.34)	-0.02	547 (1.12)	0.54	0.96
	Hub-ensemble-with-renewal	332 (1.04)	363 (1.11)	168 (0.98)	86.1 (0.93)	78.2 (1.42)	-0.02	528 (1.08)	0.58	0.96
4 wk ahead	Hub-ensemble	424 (1)	443 (1)	212 (1)	126 (1)	85.7 (1)	-0.06	675 (1)	0.58	0.92
	Hub-ensemble-realised	445 (1.05)	532 (1.2)	193 (0.91)	107 (0.85)	144 (1.68)	-0.03	700 (1.04)	0.54	0.92
	Hub-ensemble-with-all	399 (0.94)	438 (0.99)	195 (0.92)	105 (0.83)	97.9 (1.14)	-0.05	692 (1.03)	0.46	1.00
	Hub-ensemble-with-convolution	384 (0.91)	387 (0.87)	196 (0.92)	122 (0.97)	65.9 (0.77)	-0.06	602 (0.89)	0.54	0.96
	Hub-ensemble-with-crowd	407 (0.96)	456 (1.03)	202 (0.95)	105 (0.83)	101 (1.18)	-0.03	669 (0.99)	0.67	0.96
	Hub-ensemble-with-renewal	457 (1.08)	527 (1.19)	208 (0.98)	129 (1.02)	121 (1.41)	-0.06	744 (1.1)	0.50	0.96

Table S6. Scores for one and two week ahead forecasts (cut to three significant digits and rounded) for the different versions of the mean ensemble. Note that scores for cases (which include the whole period from October 12th 2020 until March 1st 2021) and deaths (which include only forecasts from the 21st of December 2020 on) are computed on different subsets. Numbers in brackets show the metrics relative to the Hub mean ensemble (i.e. the mean ensemble of all other models submitted to the German and Polish Forecast Hub, excluding our contributions). WIS is the mean weighted interval score (lower values are better), WIS - sd is the standard deviation of all scores achieved by a model. Dispersion, over-prediction and under-prediction together sum up to the weighted interval score. Bias (between -1 and 1, 0 is ideal) represents the general average tendency of a model to over- or underpredict. 50% and 90%-coverage are the percentage of observed values that fell within the 50% and 90% prediction intervals of a model.

	Model	WIS	WIS - sd	dispersion	Underpred.	Overpred.	Bias	Abs. error	50%-Cov.	90%-Cov.
Cases										
1 wk ahead	Hub-ensemble-mean	8680 (1)	10300 (1)	3700 (1)	1460 (1)	3520 (1)	-0.02	13400 (1)	0.50	0.86
	Hub-ensemble-realised-mean	7600 (0.88)	8770 (0.85)	3360 (0.91)	1090 (0.75)	3140 (0.89)	0.01	11900 (0.89)	0.52	0.90
	Hub-ensemble-with-crowd-mean	8050 (0.93)	9070 (0.88)	3520 (0.95)	1410 (0.97)	3120 (0.89)	-0.02	12600 (0.94)	0.48	0.88
	Hub-ensemble-with-renewal-mean	8090 (0.93)	9780 (0.95)	3490 (0.94)	1110 (0.76)	3490 (0.99)	0.02	12700 (0.95)	0.57	0.88
2 wk ahead	Hub-ensemble-mean	19000 (1)	22100 (1)	5960 (1)	3690 (1)	9340 (1)	0.02	28800 (1)	0.33	0.79
	Hub-ensemble-realised-mean	17100 (0.9)	20600 (0.93)	5550 (0.93)	2850 (0.77)	8660 (0.93)	0.05	26000 (0.9)	0.38	0.76
	Hub-ensemble-with-crowd-mean	17600 (0.93)	20000 (0.9)	5540 (0.93)	3790 (1.03)	8230 (0.88)	0.01	26800 (0.93)	0.36	0.76
	Hub-ensemble-with-renewal-mean	18300 (0.96)	22600 (1.02)	5910 (0.99)	2640 (0.72)	9720 (1.04)	0.06	27700 (0.96)	0.38	0.76
Deaths										
1 wk ahead	Hub-ensemble-mean	229 (1)	292 (1)	101 (1)	90.4 (1)	36.7 (1)	-0.07	315 (1)	0.71	0.92
	Hub-ensemble-realised-mean	219 (0.96)	289 (0.99)	96.8 (0.96)	79.8 (0.88)	42.6 (1.16)	-0.04	297 (0.94)	0.71	0.88
	Hub-ensemble-with-all-mean	217 (0.95)	287 (0.98)	95.3 (0.94)	83.1 (0.92)	38.7 (1.05)	-0.07	300 (0.95)	0.67	0.88
	Hub-ensemble-with-convolution-mean	225 (0.98)	292 (1)	98.7 (0.98)	94.2 (1.04)	32 (0.87)	-0.09	314 (1)	0.71	0.92
	Hub-ensemble-with-crowd-mean	222 (0.97)	289 (0.99)	98 (0.97)	84.1 (0.93)	39.6 (1.08)	-0.04	295 (0.94)	0.71	0.88
	Hub-ensemble-with-renewal-mean	225 (0.98)	290 (0.99)	99.7 (0.99)	84.7 (0.94)	40.5 (1.1)	-0.05	314 (1)	0.71	0.88
2 wk ahead	Hub-ensemble-mean	256 (1)	306 (1)	138 (1)	64.5 (1)	53.2 (1)	0.04	374 (1)	0.67	0.96
	Hub-ensemble-realised-mean	270 (1.05)	338 (1.1)	136 (0.99)	65.2 (1.01)	68.1 (1.28)	0.06	413 (1.1)	0.75	0.92
	Hub-ensemble-with-all-mean	268 (1.05)	346 (1.13)	133 (0.96)	78.7 (1.22)	57.1 (1.07)	0.05	408 (1.09)	0.75	0.96
	Hub-ensemble-with-convolution-mean	259 (1.01)	322 (1.05)	133 (0.96)	81.7 (1.27)	44.4 (0.83)	0.03	380 (1.02)	0.67	0.96
	Hub-ensemble-with-crowd-mean	264 (1.03)	315 (1.03)	133 (0.96)	70.1 (1.09)	60 (1.13)	0.06	404 (1.08)	0.71	0.96
	Hub-ensemble-with-renewal-mean	264 (1.03)	332 (1.08)	141 (1.02)	60.1 (0.93)	63.1 (1.19)	0.06	390 (1.04)	0.79	0.92

Table S7. Scores for three and four week ahead forecasts (cut to three significant digits and rounded) for the different versions of the mean ensemble. Note that scores for cases (which include the whole period from October 12th 2020 until March 1st 2021) and deaths (which include only forecasts from the 21st of December 2020 on) are computed on different subsets. Numbers in brackets show the metrics relative to the Hub mean ensemble (i.e. the mean ensemble of all other models submitted to the German and Polish Forecast Hub, excluding our contributions). WIS is the mean weighted interval score (lower values are better), WIS - sd is the standard deviation of all scores achieved by a model. Dispersion, over-prediction and under-prediction together sum up to the weighted interval score. Bias (between -1 and 1, 0 is ideal) represents the general average tendency of a model to over- or underpredict. 50% and 90%-coverage are the percentage of observed values that fell within the 50% and 90% prediction intervals of a model.

	Model	WIS	WIS - sd	dispersion	Underpred.	Overpred.	Bias	Abs. error	50%-Cov.	90%-Cov.
Cases										
3 wk ahead	Hub-ensemble-mean	35600 (1)	42100 (1)	9340 (1)	7050 (1)	19200 (1)	0.03	51200 (1)	0.26	0.62
	Hub-ensemble-realised-mean	32100 (0.9)	40500 (0.96)	8830 (0.95)	4920 (0.7)	18300 (0.95)	0.07	47200 (0.92)	0.29	0.64
	Hub-ensemble-with-crowd-mean	32200 (0.9)	36700 (0.87)	8430 (0.9)	7190 (1.02)	16500 (0.86)	0.04	46900 (0.92)	0.24	0.64
	Hub-ensemble-with-renewal-mean	35200 (0.99)	46000 (1.09)	9630 (1.03)	4600 (0.65)	20900 (1.09)	0.08	51000 (1)	0.38	0.67
4 wk ahead	Hub-ensemble-mean	60300 (1)	79300 (1)	15700 (1)	10400 (1)	34100 (1)	0.04	78600 (1)	0.29	0.57
	Hub-ensemble-realised-mean	55000 (0.91)	77100 (0.97)	14600 (0.93)	6620 (0.64)	33800 (0.99)	0.11	75200 (0.96)	0.33	0.64
	Hub-ensemble-with-crowd-mean	53400 (0.89)	66600 (0.84)	13700 (0.87)	10600 (1.02)	29200 (0.86)	0.06	70400 (0.9)	0.26	0.60
	Hub-ensemble-with-renewal-mean	61700 (1.02)	89800 (1.13)	16400 (1.04)	6400 (0.62)	38900 (1.14)	0.12	82900 (1.05)	0.31	0.64
Deaths										
3 wk ahead	Hub-ensemble-mean	289 (1)	293 (1)	178 (1)	45.9 (1)	65.7 (1)	0.01	443 (1)	0.58	1.00
	Hub-ensemble-realised-mean	310 (1.07)	348 (1.19)	182 (1.02)	42 (0.92)	86.5 (1.32)	0.08	502 (1.13)	0.58	1.00
	Hub-ensemble-with-all-mean	315 (1.09)	339 (1.16)	178 (1)	62.2 (1.36)	74 (1.13)	0.07	507 (1.14)	0.62	1.00
	Hub-ensemble-with-convolution-mean	297 (1.03)	292 (1)	174 (0.98)	67.7 (1.47)	55 (0.84)	0.01	452 (1.02)	0.67	1.00
	Hub-ensemble-with-crowd-mean	294 (1.02)	299 (1.02)	172 (0.97)	48 (1.05)	74.2 (1.13)	0.03	476 (1.07)	0.58	1.00
	Hub-ensemble-with-renewal-mean	310 (1.07)	349 (1.19)	189 (1.06)	39.4 (0.86)	81.9 (1.25)	0.05	482 (1.09)	0.62	1.00
4 wk ahead	Hub-ensemble-mean	437 (1)	568 (1)	232 (1)	72 (1)	134 (1)	0.00	702 (1)	0.62	1.00
	Hub-ensemble-realised-mean	445 (1.02)	598 (1.05)	237 (1.02)	56.4 (0.78)	152 (1.13)	0.06	707 (1.01)	0.58	1.00
	Hub-ensemble-with-all-mean	421 (0.96)	520 (0.92)	239 (1.03)	49.9 (0.69)	132 (0.99)	0.05	678 (0.97)	0.58	1.00
	Hub-ensemble-with-convolution-mean	398 (0.91)	465 (0.82)	235 (1.01)	55.6 (0.77)	107 (0.8)	0.00	628 (0.89)	0.67	1.00
	Hub-ensemble-with-crowd-mean	418 (0.96)	533 (0.94)	222 (0.96)	66.8 (0.93)	129 (0.96)	0.03	662 (0.94)	0.58	1.00
	Hub-ensemble-with-renewal-mean	467 (1.07)	636 (1.12)	248 (1.07)	61 (0.85)	158 (1.18)	0.05	755 (1.08)	0.67	1.00

Sensitivity analysis

In the original analysis, cases and deaths were scored on different periods, as the convolution model was only added later. This sensitivity shows performance of all models restricted to the period from December 14 2020 until March 1st 2021 where all models were available.

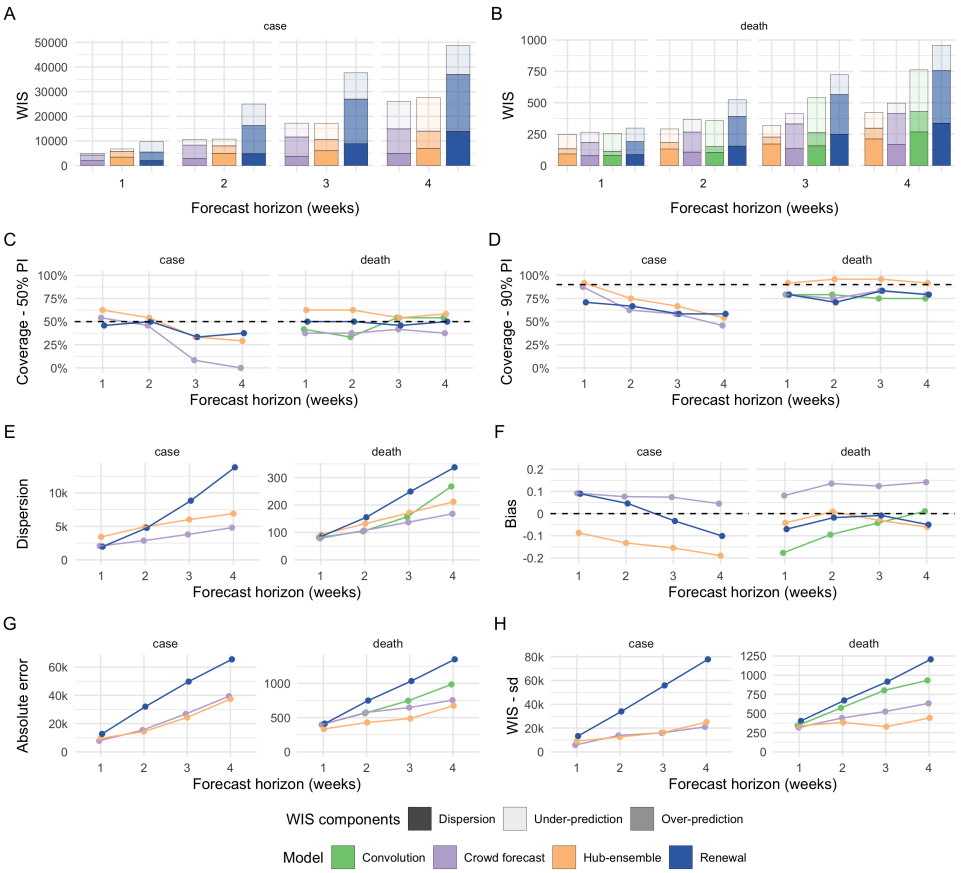

Fig S22. Visualisation of aggregate performance metrics across forecast horizons only for the period from December 14th 2020 on where all models were available. A, B: mean weighted interval score (WIS, lower indicates better performance) across horizons. WIS is decomposed into its components dispersion, over-prediction and under-prediction. C: Empirical coverage of the 50% prediction intervals (50% coverage is perfect). D: Empirical coverage of the 90% prediction intervals. E: Dispersion (same as in panel A, B). Higher values mean greater dispersion of the forecast and imply ceteris paribus a worse score. F: Bias, i.e. general (relative) tendency to over- or underpredict. Values are between -1 (complete under-prediction) and 1 (complete over-prediction) and 0 ideally. G: Absolute error of the median forecast (lower is better). H. Standard deviation of all WIS values for different horizons

Table S8. Scores for one and two week ahead forecasts (cut to three significant digits and rounded) calculated on forecasts made between December 14th 2020 and March 1st 2021. Numbers in brackets show the metrics relative to the Hub ensemble (i.e. the median ensemble of all other models submitted to the German and Polish Forecast Hub, excluding our contributions). WIS is the mean weighted interval score (lower values are better), WIS - sd is the standard deviation of all scores achieved by a model. Dispersion, over-prediction and under-prediction together sum up to the weighted interval score. Bias (between -1 and 1, 0 is ideal) represents the general average tendency of a model to over- or underpredict. 50% and 90%-coverage are the percentage of observed values that fell within the 50% and 90% prediction intervals of a model.

	Model	WIS	WIS - sd	dispersion	Underpred.	Overpred.	Bias	Abs. error	50%-Cov.	90%-Cov.
Cases										
1 wk ahead	Crowd forecast	4980 (0.74)	5730 (0.64)	2070 (0.6)	728 (0.74)	2190 (0.94)	0.09	7810 (0.82)	0.54	0.88
	Hub-ensemble	6730 (1)	8960 (1)	3430 (1)	978 (1)	2330 (1)	-0.09	9550 (1)	0.62	0.92
	Renewal	9640 (1.43)	13300 (1.48)	1970 (0.57)	4170 (4.26)	3500 (1.5)	0.09	12700 (1.33)	0.46	0.71
2 wk ahead	Crowd forecast	10700 (0.99)	13800 (1.1)	2880 (0.58)	2350 (0.85)	5430 (1.79)	0.08	15400 (1.07)	0.46	0.62
	Hub-ensemble	10800 (1)	12500 (1)	4940 (1)	2780 (1)	3030 (1)	-0.13	14400 (1)	0.54	0.75
	Renewal	25000 (2.31)	34000 (2.72)	4780 (0.97)	8710 (3.13)	11500 (3.8)	0.05	32000 (2.22)	0.50	0.67
Deaths										
1 wk ahead	Convolution	255 (1.03)	343 (1.01)	82 (0.89)	142 (1.23)	31.1 (0.75)	-0.18	399 (1.19)	0.42	0.79
	Crowd forecast	265 (1.07)	317 (0.94)	78.2 (0.85)	82 (0.71)	105 (2.52)	0.08	402 (1.2)	0.38	0.79
	Hub-ensemble	248 (1)	338 (1)	92.2 (1)	115 (1)	41.6 (1)	-0.04	334 (1)	0.62	0.92
	Renewal	298 (1.2)	403 (1.19)	87 (0.94)	107 (0.93)	105 (2.52)	-0.07	413 (1.24)	0.50	0.79
2 wk ahead	Convolution	357 (1.22)	573 (1.49)	104 (0.79)	204 (1.89)	48.8 (0.94)	-0.10	565 (1.32)	0.33	0.79
	Crowd forecast	368 (1.26)	442 (1.15)	107 (0.81)	102 (0.94)	160 (3.08)	0.14	576 (1.34)	0.38	0.75
	Hub-ensemble	292 (1)	385 (1)	132 (1)	108 (1)	51.9 (1)	0.01	429 (1)	0.62	0.96
	Renewal	524 (1.79)	671 (1.74)	155 (1.17)	133 (1.23)	236 (4.55)	-0.02	750 (1.75)	0.50	0.71

Table S9. Scores for three and four week ahead forecasts (cut to three significant digits and rounded) calculated on forecasts made between December 14th 2020 and March 1st 2021. Numbers in brackets show the metrics relative to the Hub ensemble (i.e. the median ensemble of all other models submitted to the German and Polish Forecast Hub, excluding our contributions). WIS is the mean weighted interval score (lower values are better), WIS - sd is the standard deviation of all scores achieved by a model. Dispersion, over-prediction and under-prediction together sum up to the weighted interval score. Bias (between -1 and 1, 0 is ideal) represents the general average tendency of a model to over- or underpredict. 50% and 90%-coverage are the percentage of observed values that fell within the 50% and 90% prediction intervals of a model.

	Model	WIS	WIS - sd	dispersion	Underpred.	Overpred.	Bias	Abs. error	50%-Cov.	90%-Cov.
Cases										
3 wk ahead	Crowd forecast	17200 (1)	16000 (0.98)	3800 (0.63)	5660 (0.85)	7770 (1.74)	0.07	26800 (1.1)	0.08	0.58
	Hub-ensemble	17200 (1)	16300 (1)	6030 (1)	6670 (1)	4470 (1)	-0.16	24400 (1)	0.33	0.67
	Renewal	37700 (2.19)	55900 (3.43)	8840 (1.47)	10700 (1.6)	18100 (4.05)	-0.03	49800 (2.04)	0.33	0.58
4 wk ahead	Crowd forecast	26100 (0.95)	21000 (0.84)	4810 (0.7)	11300 (0.83)	10100 (1.43)	0.04	39400 (1.05)	0.00	0.46
	Hub-ensemble	27600 (1)	25000 (1)	6900 (1)	13600 (1)	7060 (1)	-0.19	37400 (1)	0.29	0.54
	Renewal	48900 (1.77)	77800 (3.11)	13800 (2)	11900 (0.88)	23200 (3.29)	-0.10	65500 (1.75)	0.38	0.58
Deaths										
3 wk ahead	Convolution	541 (1.7)	802 (2.45)	157 (0.91)	279 (3.01)	105 (1.91)	-0.04	747 (1.53)	0.54	0.75
	Crowd forecast	414 (1.3)	526 (1.6)	137 (0.8)	82 (0.88)	194 (3.52)	0.12	648 (1.33)	0.42	0.83
	Hub-ensemble	319 (1)	328 (1)	172 (1)	92.7 (1)	55.1 (1)	-0.03	488 (1)	0.54	0.96
	Renewal	724 (2.27)	916 (2.79)	249 (1.45)	158 (1.7)	317 (5.75)	-0.01	1040 (2.13)	0.46	0.83
4 wk ahead	Convolution	763 (1.8)	932 (2.1)	268 (1.26)	331 (2.63)	164 (1.91)	0.01	985 (1.46)	0.54	0.75
	Crowd forecast	498 (1.17)	633 (1.43)	168 (0.79)	83.6 (0.66)	246 (2.87)	0.14	756 (1.12)	0.38	0.79
	Hub-ensemble	424 (1)	443 (1)	212 (1)	126 (1)	85.7 (1)	-0.06	675 (1)	0.58	0.92
	Renewal	959 (2.26)	1210 (2.73)	337 (1.59)	200 (1.59)	421 (4.91)	-0.05	1350 (2)	0.50	0.79

Overview of models and forecasters

Table S10. Overview of the models and ensembles used.

Name	Explanation
Hub-ensemble-realised	Official Forecast Hub median ensemble. Created by the Forecast Hub officially under the name 'KITCOVIDhub-median_ensemble' and used as the default ensemble. Included are our crowd forecasts as well as the renewal model (with one missed submission on December 28 2020, but not the convolution model which was deemed to similar to the renewal model.
Hub-ensemble-realised-mean	Official Forecast Hub mean ensemble. Created by the Forecast Hub officially under the name 'KITCOVIDhub-mean_ensemble'.
Hub-ensemble	Version of the official Hub median ensemble which excludes all our contributions.
Hub-ensemble-mean	Version of the official Hub mean ensemble which excludes all our contributions.
Hub-ensemble-with-renewal, Hub-ensemble-with-renewal-mean	Versions of the official Hub ensembles which of our contributions includes only the Renewal model.
Hub-ensemble-with-crowd, Hub-ensemble-with-crowd-mean	Versions of the official Hub ensembles which of our contributions includes only the Crowd forecast.
Hub-ensemble-with-convolution, Hub-ensemble-with-convolution-mean	Versions of the official Hub ensembles which of our contributions includes only the Convolution model (which originally was never included in any official Hub ensemble).
Hub-ensemble-with-all, Hub-ensemble-with-all-mean	Versions of the official Hub ensembles which includes all our contributions. For cases, this is identical to the official Hub ensembles, but for deaths the convolution model was added.
Crowd forecast	Submitted to the Forecast Hub as 'epiforecasts-EpiExpert'
Renewal model	Submitted to the Forecast Hub as 'epiforecasts-EpiNow2'
Convolution model	Submitted to the Forecast Hub as 'epiforecasts-EpiNow2_secondary'

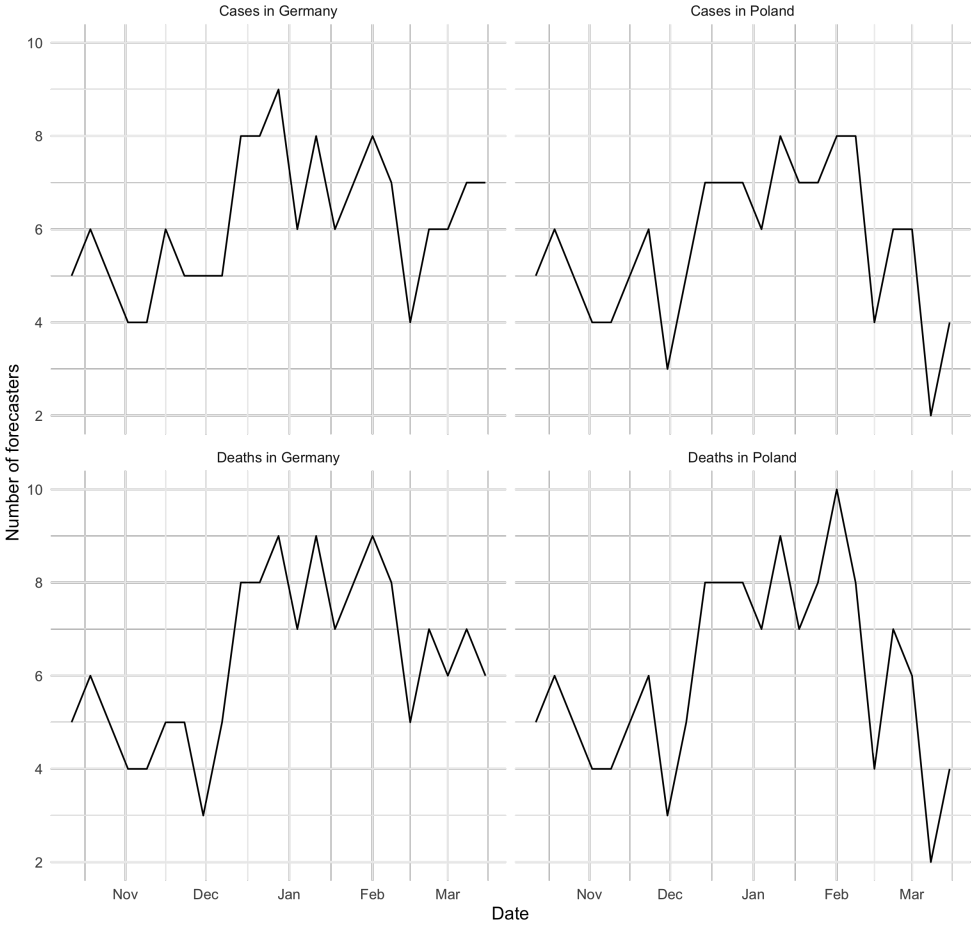

Fig S23. Number of participants who submitted a forecast over time.

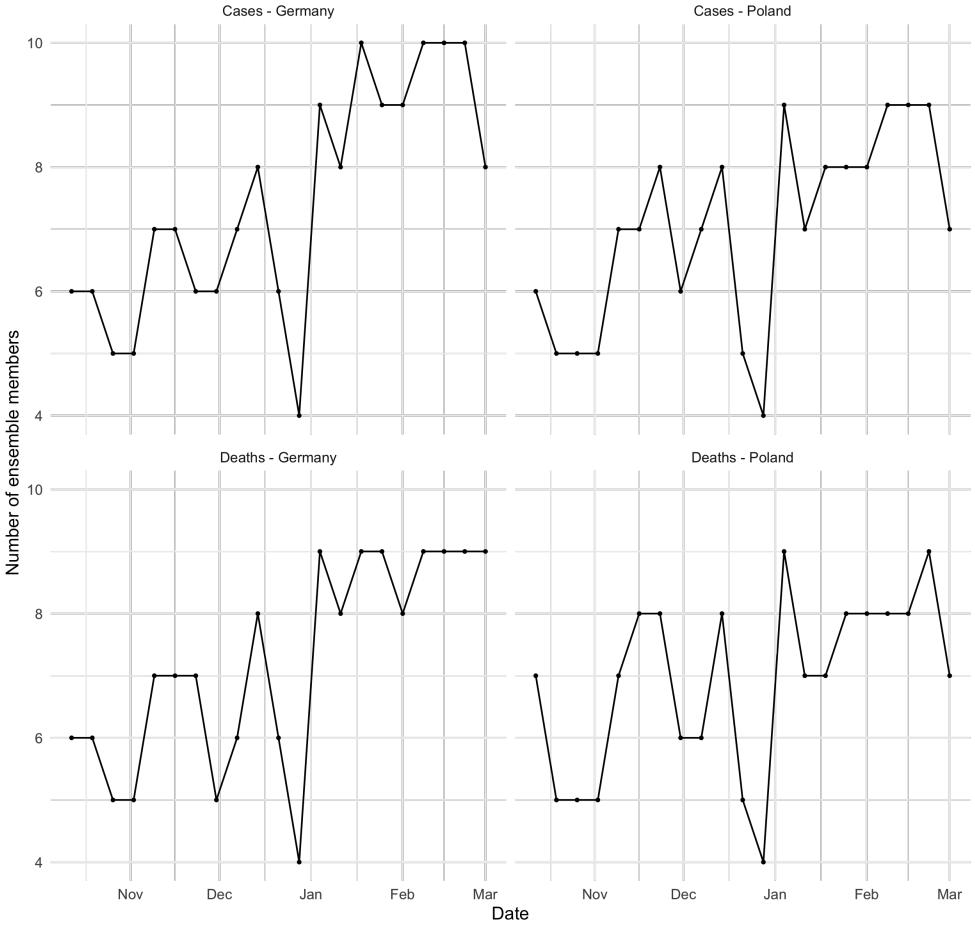

Fig S24. Number of member models (including our crowd forecasts and the renewal model) in the official Hub ensemble. Note that the renewal model was not included in the ensemble on December 28th 2020.

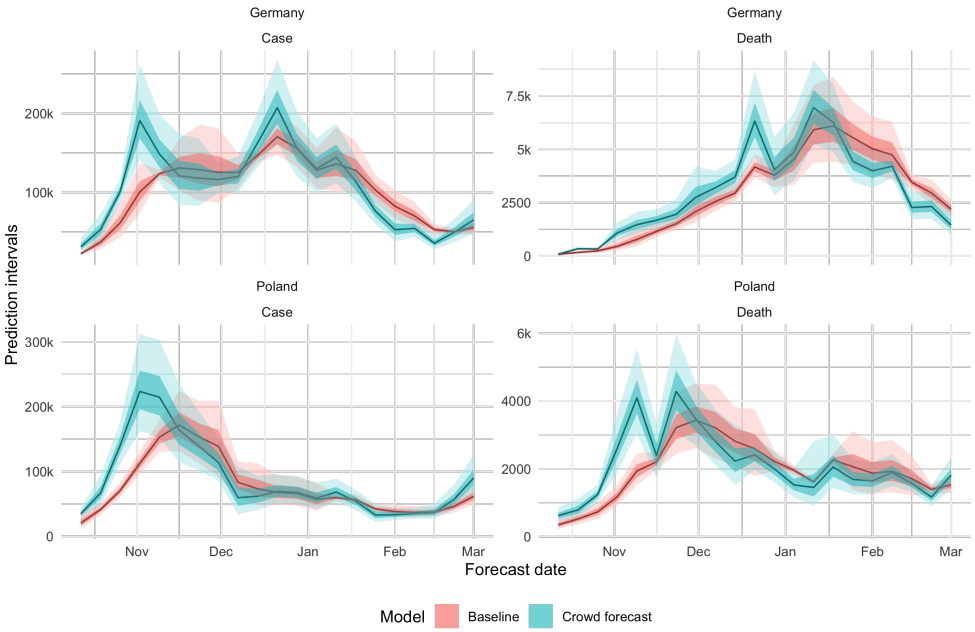

Fig S25. Crowd forecasts and baseline shown in the application for a two week horizon. Shown are the median, as well as the 50% and 90% prediction intervals (in order of decreasing opacity). For any given point in time, the baseline shown in red is what forecasters saw when they opened the app (the baseline shown was constant across all forecast horizons).

5 Transformation of forecasts for evaluating predictive performance in an epidemiological context

Chapter 4 revealed a few shortcomings of the weighted interval score as it is commonly applied in epidemiology. We saw that average scores were dominated by outliers, and overall scores scaled with incidences. This made it difficult to compare forecasts across time and location, and in particular across forecast targets. More importantly, there is a tension when evaluating forecasts on the absolute distance between forecast and observation, while the underlying process we try to model is exponential in nature. It might therefore be more appropriate to evaluate forecasts of infectious disease based on how well they capture the exponential growth rate of a disease process. This could provide a more meaningful signal about which forecasters to trust in the future, as it more closely represents the actual modelling task one has to solve in order to create an accurate representation of the spread of the infectious disease.

Different scores emphasise different kinds of errors differently. For example, in Chapter 4, the WIS penalised forecasts that overshot after missing a peak strongly. It did, however, penalise models only minimally if they were too late to predict an increase in numbers while incidences were still low. Arguably, this kind of early warning is something that policy makers would care about, but which is neglected in current evaluations. Forecast evaluations play an important role not only as a signal to modellers who aim to improve their models. They also help decision makers select which models should inform their policies in the future. If the score does not accurately reflect what policy makers care about in a good forecast, then policy makers may not pick the best model to guide their decisions.

Chapter 5 explores ways in which the forecast evaluation can be aligned more closely with what forecast consumers actually care about. One possible solution is to transform forecasts and observations before applying the WIS. We propose and analyse the natural logarithm as a transformation that is particularly attractive in an epidemiological context, but there are many more possible transformations. In particular, the idea of transforming forecasts opens up the possibility of creating composite scores, in which a score is constructed as a linear combination of scores obtained after various different transformations. This could, in the future, enable policy makers to create their own custom scores which exactly reflect what they care about.

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Scoring epidemiological forecasts on transformed scales

Nikos I. Bosse[1,2,3]*, **Sam Abbott**[1,2], **Anne Cori**[4], **Edwin van Leeuwen**[1,3,5], **Johannes Bracher**[6,7☯], **Sebastian Funk**[1,2,3☯]

1 Department of Infectious Disease Epidemiology, London School of Hygiene & Tropical Medicine, London, United Kingdom, **2** Centre for the Mathematical Modelling of Infectious Diseases, London, United Kingdom, **3** NIHR Health Protection Research Unit in Modelling & Health Economics, **4** MRC Centre for Outbreak Analysis and Modelling, Department of Infectious Disease Epidemiology, School of Public Health, Imperial College London, London, United Kingdom, **5** Modelling & Economics Unit and NIHR Health Protection Research Unit in Modelling & Health Economics, UK Health Security Agency, London, United Kingdom, **6** Chair of Statistical Methods and Econometrics, Karlsruhe Institute of Technology, Karlsruhe, Germany, **7** Computational Statistics Group, Heidelberg Institute for Theoretical Studies, Heidelberg, Germany

☯ These authors contributed equally to this work.
* nikos.bosse@lshtm.ac.uk

Citation: Bosse NI, Abbott S, Cori A, van Leeuwen E, Bracher J, Funk S (2023) Scoring epidemiological forecasts on transformed scales. PLoS Comput Biol 19(8): e1011393. https://doi.org/10.1371/journal.pcbi.1011393

Editor: James M McCaw, The University of Melbourne, AUSTRALIA

Received: January 30, 2023

Accepted: July 27, 2023

Published: August 29, 2023

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcbi.1011393

Data Availability Statement: All code and data is available at https://github.com/epiforecasts/transformation-forecast-evaluation.

Funding: NIB received funding from the National Institute for Health and Care Research (NIHR)

Abstract

Forecast evaluation is essential for the development of predictive epidemic models and can inform their use for public health decision-making. Common scores to evaluate epidemiological forecasts are the Continuous Ranked Probability Score (CRPS) and the Weighted Interval Score (WIS), which can be seen as measures of the absolute distance between the forecast distribution and the observation. However, applying these scores directly to predicted and observed incidence counts may not be the most appropriate due to the exponential nature of epidemic processes and the varying magnitudes of observed values across space and time. In this paper, we argue that transforming counts before applying scores such as the CRPS or WIS can effectively mitigate these difficulties and yield epidemiologically meaningful and easily interpretable results. Using the CRPS on log-transformed values as an example, we list three attractive properties: Firstly, it can be interpreted as a probabilistic version of a relative error. Secondly, it reflects how well models predicted the time-varying epidemic growth rate. And lastly, using arguments on variance-stabilizing transformations, it can be shown that under the assumption of a quadratic mean-variance relationship, the logarithmic transformation leads to expected CRPS values which are independent of the order of magnitude of the predicted quantity. Applying a transformation of log (x + 1) to data and forecasts from the European COVID-19 Forecast Hub, we find that it changes model rankings regardless of stratification by forecast date, location or target types. Situations in which models missed the beginning of upward swings are more strongly emphasised while failing to predict a downturn following a peak is less severely penalised when scoring transformed forecasts as opposed to untransformed ones. We conclude that appropriate transformations, of which the natural logarithm is only one particularly attractive option, should be considered when assessing the performance of different models in the context of infectious disease incidence.

Health Protection Research Unit (HPRU) in Modelling and Health Economics (grant code NIHR200908). SA's work was funded by the Wellcome Trust (grant: 210758/Z/18/Z). AC acknowledges funding from the MRC Centre for Global Infectious Disease Analysis (reference MR/R015600/1) jointly funded by the UK Medical Research Council (MRC) and the UK Foreign, Commonwealth & Development Office (FCDO), under the MRC/FCDO Concordat agreement and is also part of the EDCTP2 programme supported by the European Union; the Academy of Medical Sciences Springboard, funded by the Academy of Medical Sciences, Wellcome Trust, the Department for Business, Energy and Industrial Strategy, the British Heart Foundation, and Diabetes UK (reference SBF005\1044); and the National Institute for Health and Care Research (NIHR) Health Protection Research Unit in Modelling and Health Economics, a partnership between the UK Health Security Agency, Imperial College London and LSHTM (grant code NIHR200908). EvL acknowledges funding by the National Institute for Health and Care Research (NIHR) Health Protection Research Unit (HPRU) in Modelling and Health Economics (grant number NIHR200908) and the European Union's Horizon 2020 research and innovation programme - project EpiPose (101003688). The work of JB was supported by the Helmholtz Information and Data Science Project SIMCARD as well as Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 512483310. SF's work was supported by the Wellcome Trust (grant: 210758/Z/18/Z) and the HPRU (grant code NIHR200908). The views expressed are those of the authors and not necessarily those of the UK Department of Health and Social Care (DHSC), NIHR, or UKHSA. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Author summary

Scores like the Continuous Ranked Probability Score (CRPS) or the Weighted Interval Score (WIS) are commonly used to evaluate epidemiological forecasts and are a measure of absolute distance between forecast and observation. Due to the exponential nature of epidemic processes, evaluating the absolute distance between forecast and observation may not be ideal. We argue that transforming counts before applying the CRPS or WIS can yield more meaningful results. The natural logarithm is a particularly attractive transformation in epidemiological settings. Scores computed on log-transformed values can be interpreted as a probabilistic version of a relative error and reflect how well forecasters predict the time-varying epidemic growth rate. If the data-generating process has a quadratic mean-variance relationship, the logarithmic transformation also leads to expected CRPS values which are independent of the order of magnitude of the predicted quantity. We illustrate these properties using data from the European COVID-19 Forecast Hub and find that scoring transformed counts changes model rankings. Stronger emphasis is given to situations in which forecasters missed the beginning of upward swings, while failing to predict a downturn following a peak is less severely penalised. We generally recommend including evaluations of transformed counts when assessing forecaster performance.

Introduction

Probabilistic forecasts [1] play an important role in decision-making in epidemiology and public health [2], as well as other areas as diverse as economics [3] or meteorology [4]. Forecasts based on epidemiological modelling in particular have received widespread attention during the COVID-19 pandemic. Evaluations of forecasts can provide feedback for researchers to improve their models and train ensembles. They moreover help decision-makers distinguish good from bad predictions and choose forecasters and models that are best suited to inform future decisions.

Probabilistic forecasts are usually evaluated using so-called (strictly) proper scoring rules [5], which return a numerical score as a function of the forecast and the observed data. Proper scoring rules are constructed such that they encourage honest forecasting and cannot be 'gamed' or 'cheated'. Assuming that the forecaster's actual best judgement corresponds to a predictive distribution F, a proper score is constructed such that if F was the data-generating process, no other distribution G would yield a better expected score. A scoring rule is called *strictly* proper if there is no other distribution that under F achieves the *same* expected score as F, meaning that any deviation from F leads to a worsening of expected scores. Forecasters (anyone or anything that issues a forecast) are thus incentivised to report their true belief F about the future. Common proper scoring rules are the logarithmic or log score [6] and the continuous ranked probability score (CRPS, [5]). The log score is the predictive log density or probability mass evaluated at the observed value. It is supported by the likelihood principle [7] and has many desirable theoretical properties; however, the particularly severe penalties it assigns to occasional misguided forecasts make it little robust [8]. Moreover, it is not easily applied to forecasts reported as samples or quantiles, as used in many recent disease forecasting efforts. It is nonetheless occasionally used in epidemiology (see e.g., [1, 9]), but in recent years the CRPS and the weighted interval score (WIS, [8]) have become increasingly popular.

The CRPS measures the distance of the predictive distribution to the observed data as

$$\mathrm{CRPS}(F, y) = \int_{-\infty}^{\infty} \left(F(x) - \mathbf{1}(x \geq y)\right)^2 dx, \tag{1}$$

where y is the true observed value, F is the cumulative distribution function (CDF) of the predictive distribution, and $\mathbf{1}()$ is the indicator function. The CRPS can be understood as a generalisation of the absolute error to predictive distributions, and interpreted on the natural scale of the data. The WIS is an approximation of the CRPS for predictive distributions represented by a set of predictive quantiles and is currently used to assess forecasts in the so-called COVID-19 Forecast Hubs in the US [10, 11], Europe [12] and Germany and Poland [13, 14], as well as the US *FluSight* project on influenza forecasting [15]. The WIS is defined as

$$\mathrm{WIS}(F, y) = \frac{1}{K} \times \sum_{k=1}^{K} 2 \times \left[\mathbf{1}(y \leq q_{\tau_k}) - \tau_k\right] \times (q_{\tau_k} - y), \tag{2}$$

where q_τ is the τ quantile of the forecast F, y is the observed outcome and K is the number of (roughly equally spaced) predictive quantiles provided. The WIS can be decomposed into three components, dispersion, underprediction and overprediction, which reflect the spread of the forecast and whether it was centred above or below the observed value. We show an alternative definition based on central prediction intervals in S1 Text which illustrates this decomposition.

The notion of absolute distance encoded by the CRPS and WIS provides a straightforward interpretation, but may not always be the most useful perspective in the context of infectious disease spread. Especially in their early phase, outbreaks are best conceived as exponential processes, characterized by potentially time varying reproduction numbers R_t [16] or epidemic growth rates r_t [17]. If the true modelling task revolves around estimating and forecasting these quantities, then evaluating forecasts based on the absolute distance between forecasted and observed incidence values penalises underprediction (of the reproduction number or growth rate) less than overprediction by the same amount. For illustration, consider an incidence forecast issued at time 0 and referring to time t that misses the correct average growth rate $\bar{r}_t$ by either $-\epsilon$ or $+\epsilon$. Then the ratio of the resulting absolute errors on the scale of observed incidences y_t is

$$\frac{|y_0 \exp[(\bar{r}_t - \epsilon) \times t] - y_0 \exp(\bar{r}_t t)|}{|y_0 \exp[(\bar{r}_t + \epsilon) \times t] - y_0 \exp(\bar{r}_t t)|} = \exp(-\epsilon t) < 1. \tag{3}$$

If one is to measure the ability to forecast the underlying infection dynamics, it may thus be more desirable to evaluate errors on the scale of the growth rate directly.

Another argument against using notions of absolute distance between predicted and observed incidence values is that forecast consumers may find errors on a relative scale easier to interpret and more useful in order to track predictive performance across targets of different orders of magnitude. [18] have proposed the scaled CRPS (SCRPS) which is locally scale invariant; however, it does not correspond to a relative error measure and lacks a straightforward interpretation as available for the CRPS.

Lastly, it may be considered desirable to give all forecast targets similar weight in an overall performance evaluation. As the CRPS typically scales with the order of magnitude of the quantity to be predicted, this is not the case for the CRPS, which will typically assign higher scores to forecast targets with high expected values (e.g., in large locations or around the peak of an epidemic). Bracher et al. [8] have argued that this is a desirable feature, directing attention to situations of particular public health relevance. An evaluation based on absolute errors,

however, will assign little weight to other potentially important aspects, such as the ability to correctly predict future upswings while observed numbers are still low.

In many fields, it is common practice to forecast transformed quantities (see e.g. [19] in finance, [20] in macroeconomics, [21] in hydrology or [22] in meteorology). While the goal of the transformations is often to improve the accuracy of the predictions, they can also be used to enhance and complement the evaluation process. In this paper, we argue that the aforementioned issues with evaluating epidemic forecasts based on measures of absolute error on the natural scale can be addressed by transforming the forecasts and observations prior to scoring using some strictly monotonic transformation. Strictly monotonic transformations can shift the focus of the evaluation in a way that may be more appropriate for epidemiological forecasts, while guaranteeing that the score remains proper. Many different transformations may be appropriate and useful, depending on the exact context, the desired focus of the evaluation, and specific aspects forecast consumers care most about (see Discussion).

For conceptual clarity and to allow for a more in-depth discussion, we focus mostly on the natural logarithm as a particularly attractive transformation in the context of epidemic phenomena. We refer to this transformation as 'log-transformation' and to scores that have been computed from log-transformed forecasts and observations as scores 'on the log scale' (as opposed to scores 'on the natural scale', which involve no transformation). In the theoretical part of the paper, 'log-transformation' and 'log scale' generally refer to a transformation of $\log_e(x)$. For practical applications in the later sections we also use these terms to describe a transformation of $\log_e(x + a)$ with a small $a > 0$ in order to keep the terminology and notation simple. For a prediction target with strictly positive support, the CRPS after applying a log-transformation is given by

$$\mathrm{CRPS}(F_{\log}, \log y) = \int_{-\infty}^{\infty} (F_{\log}(x) - \mathbf{1}(x \geq \log y))^2 dx. \tag{4}$$

Here, y is again the observed outcome and $F_{\log}$ is the predictive CDF of the log-transformed outcome, i.e.,

$$F_{\log}(x) = F(\exp(x)), \tag{5}$$

with F the CDF on the original scale. Instead of a score representing the magnitude of absolute errors, applying a log-transformation prior to the CRPS yields a score which a) measures relative error, b) provides a measure for how well a forecast captures the exponential growth rate of the target quantity and c) is less dependent on the expected order of magnitude of the quantity to be predicted). We therefore argue that such evaluations on the logarithmic scale should complement the prevailing evaluations on the natural scale. Other transformations may likewise be of interest. We briefly explore the square root transformation as an alternative transformation. Our analysis mostly focuses on the CRPS (or WIS) as an evaluation metric for probabilistic forecasts, given its widespread use throughout the COVID-19 pandemic. We note that the logarithmic score has scale invariance properties which imply that score differences between different forecasts are invariant to strictly monotonic transformations (see [23] on corresponding properties of likelihood ratios and [24]). The question of the right scale to evaluate forecasts on does therefore not arise for the log score.

The remainder of the article is structured as follows. First, we provide some mathematical intuition on applying the log-transformation prior to evaluating the CRPS, highlighting the connections to relative error measures, the epidemic growth rate and variance stabilizing transformations. We then discuss the effect of the log-transformation on forecast rankings as well as practical considerations for applying transformations in general and the log-

transformation in particular. To analyse the real-world implications of the log-transformation we use forecasts submitted to the European COVID-19 Forecast Hub [12, 25]. Finally, we provide scoring recommendations, discuss alternative transformations that may be useful in different contexts, and suggest further research avenues).

Logarithmic transformation of forecasts and observations

Interpretation as a relative error

To illustrate the effect of applying the natural logarithm prior to evaluating forecasts we consider the absolute error, which the CRPS and WIS generalize to probabilistic forecasts. We assume strictly positive support (meaning that no specific handling of zero values is needed), a restriction we will address when applying this transformation in practice. When considering a point forecast $\hat{y}$ for a quantity of interest y, such that

$$y = \hat{y} + \varepsilon, \tag{6}$$

the absolute error is given by $|\varepsilon|$. When taking the logarithm of the forecast and the observation first, thus considering

$$\log y = \log \hat{y} + \varepsilon^*, \tag{7}$$

the resulting absolute error $|\varepsilon^*|$ can be interpreted as an approximation of various common relative error measures. Using that $\log(a) \approx a - 1$ if a is close to 1, we get

$$|\varepsilon^*| = |\log \hat{y} - \log y| = \left|\log\left(\frac{\hat{y}}{y}\right)\right| \overset{\text{if } \hat{y} \approx y}{\approx} \left|\frac{\hat{y}}{y} - 1\right| = \left|\frac{\hat{y} - y}{y}\right|. \tag{8}$$

The absolute error after log transforming is thus an approximation of the *absolute percentage error* (APE, [26]) as long as forecast and observation are close. As we assumed that $\hat{y} \approx y$, we can also interpret it as an approximation of the *relative error* (RE, [26])

$$\left|\frac{\hat{y} - y}{\hat{y}}\right| \tag{9}$$

and the *symmetric absolute percentage error* (SAPE; see e.g., [27])

$$\left|\frac{\hat{y} - y}{y/2 + \hat{y}/2}\right|. \tag{10}$$

As Fig 1 shows, the alignment with the SAPE is in fact the closest and holds quite well even if predicted and observed value differ by a factor of two or three. Generalising to probabilistic forecasts, the CRPS applied to log-transformed forecasts and outcomes can thus be seen as a probabilistic counterpart to the symmetric absolute percentage error, which offers an appealing intuitive interpretation.

Interpretation as scoring the exponential growth rate

Another interpretation for the log-transform is possible if the generative process is framed as exponential with a time-varying growth rate $r(t)$ (see e.g. [28]), i.e.

$$\frac{d}{dt}y(t) = r(t)y(t) \tag{11}$$

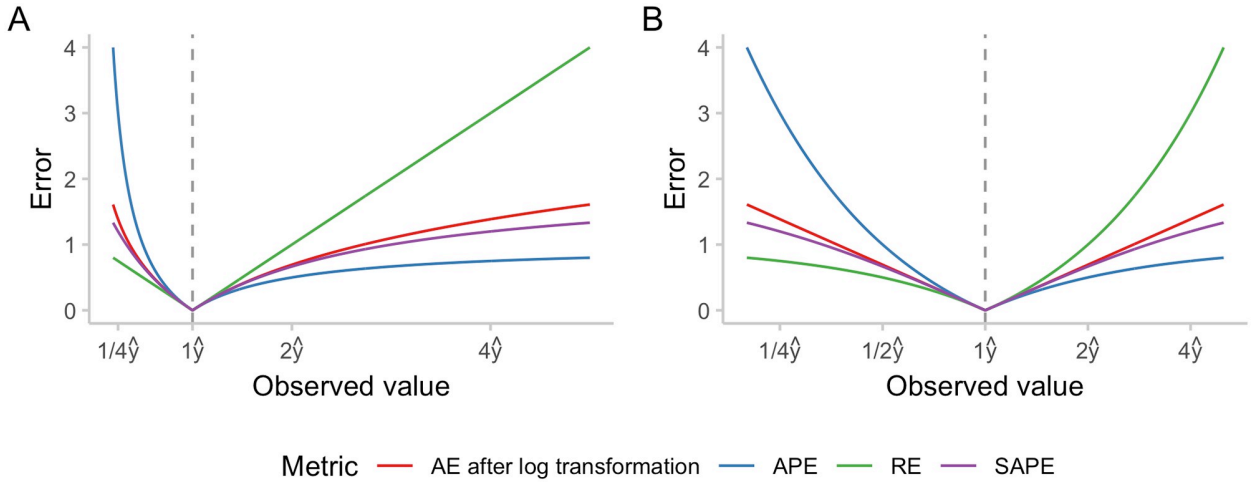

Fig 1. Numerical comparison of different measures of relative error: Absolute percentage error (APE), relative error (RE), symmetric absolute percentage error (SAPE) and the absolute error applied to log-transformed predictions and observations. We denote the predicted value by $\hat{y}$ and display errors as a function of the ratio of observed and predicted value. A: x-axis shown on a linear scale. B: x-axis shown on a logarithmic scale.

https://doi.org/10.1371/journal.pcbi.1011393.g001

which is solved by

$$y(t) = y_0 \exp\left(\int_0^t r(t')dt'\right) = y_0 \exp(\bar{r}_t t) \tag{12}$$

where y_0 is an initial data point and $\bar{r}_t$ is the mean of the growth rate between the initial time point 0 and time t.

If a forecast $\hat{y}(t)$ for the value of the time series at time t is issued at time 0 based on the data point y_0 then the absolute error after log transformation is

$$\begin{aligned} \epsilon^* &= |\log[\hat{y}(t)] - \log[y(t)]| \\ &= |\log[y_0 \exp(\hat{\bar{r}}_t t)] - \log[y_0 \exp(\bar{r}_t t)]| \\ &= t|\hat{\bar{r}}_t - \bar{r}_t| \end{aligned} \tag{13}$$

where $\bar{r}_t$ is the true mean growth rate and $\hat{\bar{r}}_t$ is the forecast mean growth rate. We thus evaluate the error in the mean exponential growth rate, scaled by the length of the time period considered. Again generalising this to the CRPS and WIS implies a probabilistic evaluation of forecasts of the epidemic growth rate.

Interpretation as a variance-stabilising transformation

When evaluating models across sets of forecasting tasks, it may be desirable for each target to have a similar impact on the overall results. This could be motivated by the assumption that forecasts from different geographical units and time periods provide similar amounts of information about how well a forecaster performs. One would then like the resulting scores to be independent of the order of magnitude of the target to predict. CRPS values on the natural scale, however, typically scale with the order of magnitude of the quantity to be predicted. Average scores are then dominated by the results achieved for targets with high expected outcomes in a way that does not necessarily reflect the underlying predictive ability well.

If the predictive distribution for the quantity Y equals the true data-generating process F (an ideal forecast), the expected CRPS is given by [5]

$$\mathbb{E}[\mathrm{CRPS}(F, y)] = 0.5 \times \mathbb{E}|Y - Y'|, \tag{14}$$

where Y and Y' are independent samples from F. This corresponds to half the *mean absolute difference*, which is a measure of dispersion. If F is well-approximated by a normal distribution $\mathrm{N}(\mu, \sigma^2)$, the approximation

$$\mathbb{E}_F[\mathrm{CRPS}(F, y)] \approx \frac{\sigma}{\sqrt{\pi}} \tag{15}$$

can be used. This means that the expected CRPS scales roughly with the standard deviation, which in turn typically increases with the mean in epidemiological forecasting. In order to make the expected CRPS independent of the expected outcome, a *variance-stabilising transformation* (VST, [29, 30]) can be employed. The choice of this transformation depends on the mean-variance relationship of the underlying process.

If the mean-variance relationship of the data-generating distribution is quadratic with $\sigma^2 = c \times \mu^2$, the natural logarithm can serve as the VST. Denoting by $F_{\log}$ the predictive distribution for $\log(Y)$, we can use the delta method (a first-order Taylor approximation, see e.g., [30]), to show that

$$\mathbb{E}_F[\mathrm{CRPS}\{F_{\log}, \log(y)\}] \approx \frac{\sigma/\mu}{\sqrt{\pi}} = \frac{\sqrt{c}}{\sqrt{\pi}}. \tag{16}$$

As σ and μ are linked through the quadratic mean-variance relationship (or linear mean-standard deviation relationship, $\sigma = \sqrt{c} \times \mu$), the expected CRPS thus stays constant regardless of the expected value of the data-generating distribution μ. The assumption of a quadratic mean-variance relationship is closely linked to the aspects discussed earlier. It implies that relative errors have constant variance and can thus be meaningfully compared across different targets. Also, it arises naturally if we assume that our capacity to predict the epidemic growth rate does not depend on the expected outcome, i.e. does not depend on the current phase of the epidemic or the order of magnitude of current observations.

If the mean-variance relationship is linear with $\sigma^2 = c \times \mu$, as with a Poisson-distributed variable, the square root is known to be a VST [30]. Denoting by $F_{\sqrt{\ }}$ the predictive distribution for $\sqrt{Y}$, the delta method can again be used to show that

$$\mathbb{E}_F[\mathrm{CRPS}\{F_{\sqrt{\ }}, \sqrt{y}\}] \approx \frac{\sigma/\sqrt{\mu}}{2\sqrt{\pi}} = \frac{\sqrt{c}}{2\sqrt{\pi}}. \tag{17}$$

We note that while standard in the derivation of variance-stabilizing transformations, the application of the delta method in Eqs (16) and (17) requires the probability mass of F to be tightly distributed. If this is not the case, the approximation and thus the variance stabilization may be less accurate.

To strengthen our intuition on how transforming outcomes prior to applying the CRPS shifts the emphasis between targets with high and low expected outcomes, Fig 2 shows the expected CRPS of ideal forecasters under different mean-variance relationships and transformations. We consider a Poisson distribution where $\sigma^2 = \mu$, a negative binomial distribution with size parameter $\theta = 10$ and thus $\sigma^2 = \mu + \mu^2/10$, and a truncated normal distribution with practically constant variance. We see that when applying the CRPS on the natural scale, the expected CRPS grows monotonically as the variance of the predictive distribution (which is equal to the data-generating distribution for the ideal forecaster) increases. The expected

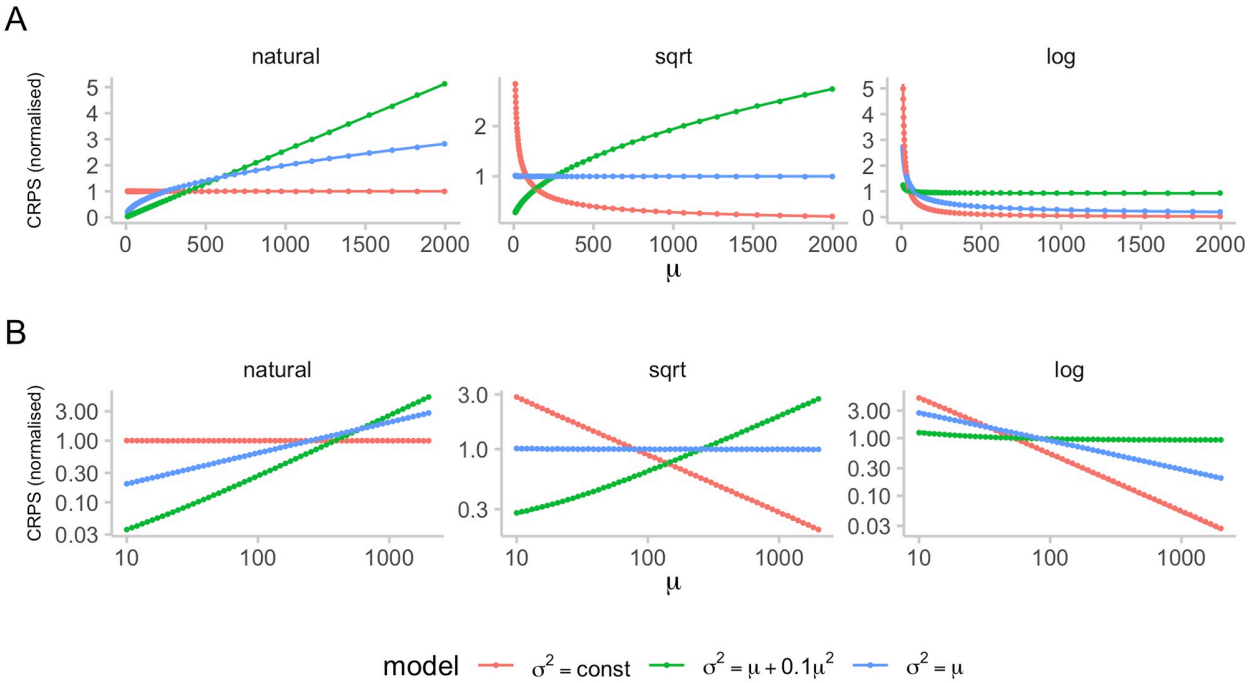

Fig 2. Expected CRPS scores as a function of the mean and variance of the forecast quantity. We computed expected CRPS values for three different distributions, assuming an ideal forecaster with predictive distribution equal to the true underlying (data-generating) distribution. These expected CRPS values were computed for different predictive means based on 10,000 samples each and are represented by dots. Solid lines show the corresponding approximations of the expected CRPS from Eqs (16) and (17). S3 Fig shows the quality of the approximation in more detail. The first distribution (red) is a truncated normal distribution with constant variance (we chose $\sigma = 1$ in order to only obtain positive samples). The second (green) is a negative binomial distribution with variance $\theta = 10$ and variance $\sigma^2 = \mu + 0.1\mu^2$. The third (blue) is a Poisson distribution with $\sigma^2 = \mu$. To make the scores for the different distributions comparable, scores were normalised to one, meaning that the mean score for every distribution (red, green, blue) is one. A: Normalised expected CRPS for ideal forecasts with increasing means for three distribution with different relationships between mean and variance. Expected CRPS was computed on the natural scale (left), after applying a square-root transformation (middle), and after adding one and applying a log-transformation to the data (right). B: A but with x and y axes on the log scale.

https://doi.org/10.1371/journal.pcbi.1011393.g002

CRPS is constant only for the distribution with constant variance, and grows in μ for the other two. When applying a log-transformation first, the expected CRPS is almost independent of μ for the negative binomial distribution and large μ, while smaller targets have higher expected CRPS in case of the Poisson distribution and the normal distribution with constant variance. When applying a square-root-transformation, the expected CRPS is independent of the mean for the Poisson-distribution, but not for the other two (with a positive relationship in the normal case and a negative one for the negative binomial). As can be seen in Fig 2 and S3 Fig, the approximations presented in Eqs (16) and (17) work quite well for our simulated example.

Effects on model rankings

Rankings between different forecasters based on the CRPS may change when making use of a transformation, both in terms of aggregate and individual scores. We illustrate this in Fig 3 with two forecasters, A and B, issuing two different distributions with different dispersion. When showing the obtained CRPS as a function of the observed value, it can be seen that the ranking between the two forecasters may change when scoring the forecast on the logarithmic, rather than the natural scale. In particular, on the natural scale, forecaster A, who issues a more uncertain distribution, receives a better score than forecaster B for observed values far away from the centre of the respective predictive distribution. On the log scale, however,

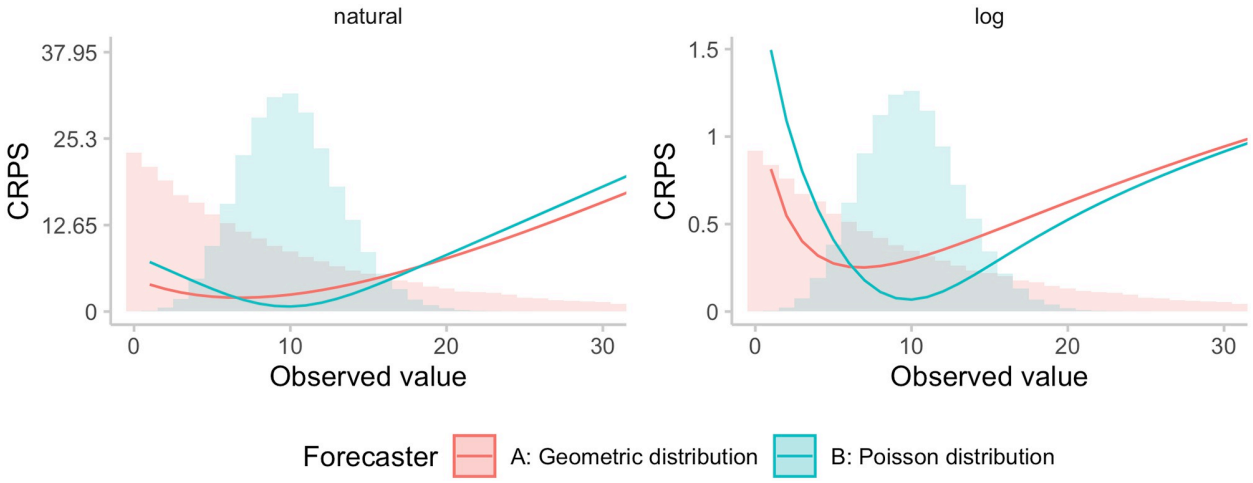

Fig 3. Illustration of the effect of the log-transformation of the ranking for a single forecast. Shown are CRPS (or WIS, respectively) values as a function of the observed value for two forecasters. Model A issues a geometric distribution (a negative binomial distribution with size parameter $\theta = 1$) with mean $\mu = 10$ and variance $\sigma^2 = \mu + \mu^2 = 110$), while Model B issues a Poisson distribution with mean and variance equal to 10. Zeroes in this illustrative example were handled by adding one before applying the natural logarithm.

https://doi.org/10.1371/journal.pcbi.1011393.g003

forecaster A receives a lower score for large observed values, being more heavily penalised for assigning large probability to small values (which, in relative terms, are far away from the actual observation). We note that the chosen example involving a geometric forecast distribution is somewhat constructed; as illustrated in our practical example, rankings between models in practice stay quite stable for a single forecast.

Overall model rankings would be expected to differ more when scores are averaged across multiple forecasts or targets. The change in rankings of aggregate scores usually is mainly driven by the order of magnitude of scores for different forecast targets across time, location and target type and less so by the kind of changes in model rankings for single forecasts discussed above. Large observations will dominate average CRPS values when evaluation is done on the natural scale, but much less so after log transformation. Depending on how different models perform across targets of different orders of magnitude, rankings in terms of average scores may change when applying a transformation.

Practical considerations and other transformations

In practice, one issue with the log transform is that it is not readily applicable to negative or zero values, which need to be removed or otherwise handled. One common approach to this end is to add a small positive quantity, such as $a = 1$, to all observations and predictions before taking the logarithm [31]. This still represents a strictly monotonic transformation, but the choice of a does influence scores and rankings (measures of relative errors shrink the larger the chosen value a). As a rule of thumb, if if $x > 5a$, the difference between $\log(x + a)$ and $\log(x)$ is small, and it becomes negligible if $x > 50a$. Choosing a suitable offset a thus balances two competing concerns: on the one hand, choosing a small a makes sure that the transformation is as close to a natural logarithm as possible and scores can be interpreted as outlined in the previous sections. On the other hand, choosing a larger a can help stabilise scores for forecasts and observations close to zero, avoiding giving excessive weight to forecasts of small quantities. For increasing a, less relative weight is given to smaller forecast targets. For very large values of a, $\log(x + a)$ is roughly linear in x, so that using a very large a implies similar relative weighting as applying no transformation at all. In practice, a user could explore the effect of different

values of a graphically and choose a such that the relative weightings of times and regions with high and low incidence correspond to their preferences.

A related issue occurs when the predictive distribution has a large probability mass on zero (or on very small values), as this can translate into an excessively wide forecast in relative terms. In our applied example this is illustrated in S7 Fig. In such instances, the dispersion component of the WIS is inflated for scores obtained after applying the natural logarithm because forecasts contained zero in its prediction intervals. To deal with this issue one could choose to use a higher a value when applying a transformation $\log(x + a)$, for example $a = 10$ instead of the $a = 1$ that we chose to use.

A natural question is which other transformations could be applied and whether resulting scores remain (strictly) proper. In principle, any transformation function can be applied simultaneously to forecasts and observations as long as the definition of the transformation is independent of the forecasts and any quantities unknown at the time of forecasting, including the observed value. This simply corresponds to a re-definition of the forecasting target. However, applying non-invertible transformations leads to a loss in information conveyed by forecasts, which we consider undesirable. The resulting score will be proper, but it may not be strictly proper anymore (as forecasts differing from the forecaster's true belief on the original scale may be identical on the transformed scale). When using the CRPS or the WIS, it seems most appropriate to use only strictly monotonic transformations such as the natural logarithm or the square root as otherwise the encoded notion of distance may become meaningless.

Some other strictly monotonic transformations that can be applied are scaling by the population size or scaling by past observations. The latter is similar to applying a log-transformation, but corresponds to evaluating a forecast of multiplicative, rather than exponential growth rates. The arising issue of dividing by zero can again be solved by adding a small offset a. Scaling a forecast by the later observed value (as opposed to scaling by past observations) is generally not permissible as it can result in improper scores (see [32] on the closely related topic of weighting scores with a function of the observed value). Similarly, scaling forecasts and observations by a function of the predictive distribution (like the predictive mean) may lead to improper scores; however, we are unaware of existing theoretical arguments on this.

When applying a transformation, the order of the operations matters, and applying a transformation after scores have been computed generally does not guarantee that the score remains proper. In the case of log transforms, taking the logarithm of the CRPS values, rather than scoring the log-transformed forecasts and data, results in an improper score. We illustrate this point using simulated data in S1 Fig, where it can be seen that in the example overconfident models perform best in terms of the log WIS. We note that strictly speaking, re-scaling average scores by the average score of a baseline model or average scores across different models to obtain skill scores likewise leads to improper scores [5]. The application of such skill scores, however, is established practice and considered largely unproblematic.

We note that in the practical evaluation of operational forecasting systems several additional challenges arise, which we do not study in detail. These concern e.g., the removal of outlying observations and forecasts and the handling of missing forecasts. The solutions we employed in practice are detailed below.

Empirical example: The European Forecast Hub

Setting

As an empirical comparison of evaluating forecasts on the natural and on the log scale, we use forecasts from the European Forecast Hub [12, 25]. The European COVID-19 Forecast Hub is one of several COVID-19 Forecast Hubs [11, 13] which have been systematically collecting,

aggregating and evaluating forecasts of several COVID-19 targets created by different teams every week. Forecasts are made one to four weeks ahead into the future and follow a quantile-based format with a set of 23 quantiles (0.01, 0.025, 0.05, . . ., 0.5, . . .0.95, 0.975, 0.99).

The forecasts used for the purpose of this illustration are forecasts submitted between the 8th of March 2021 and the 5th of December 2022 for reported cases and deaths from COVID-19. Target dates range from the 13th of March 2021 to the 10th of December 2022, for a total of 92 weeks. See [12] for a more thorough description of the data. We filtered all forecasts submitted to the Hub to only include the seven models which have submitted forecasts for both deaths and cases for 4 horizons in 32 locations on at least 46 forecast dates (see S4 Fig). We removed all observations marked as data anomalies by the European Forecast Hub [12] as well as all remaining negative observed values. These anomalies made up a relevant fraction of all observations. On average across locations, 12.1 out of 92 (13.2%) observations were removed for cases and 12.4 out of 92 (13.5%) for deaths. S5 Fig displays the number of anomalies removed for each location. In addition, we filtered out a small number of erroneous forecasts that were in extremely poor agreement with the observed data, as defined by any of the conditions listed in S2 Table. S6 Fig shows the percentage of forecasts removed for each model. Those few (less than 0.2% of forecasts for each model) erroneous outlier forecasts had excessive influence on average scores and relative skill scores in a way that was not representative of normal model behaviour. We removed them here in order to better illustrate the effects of the log-transformation on scores that one would expect in a well-behaved scenario. In a regular forecast evaluation such erroneous forecasts should usually not be removed and would count towards overall model scores.

All predictive quantiles were truncated at 0. We applied the log-transformation after adding a constant $a = 1$ to all predictions and observed values. The choice of $a = 1$ in part reflects convention, but also represents a suitable choice as it avoids giving excessive weight to forecasts close to zero, while at the same time ensuring that scores for observations >5 can be interpreted reasonably. S2 Fig illustrates the effect of adding a small quantity before taking the logarithm. The analysis was conducted in `R` [33], using the `scoringutils` package [34] for forecast evaluation. All code is available on GitHub (https://github.com/epiforecasts/transformation-forecast-evaluation). Where not otherwise stated, we report results for a two-week-ahead forecast horizon.

In addition to the WIS we use pairwise comparisons [11] to evaluate the relative performance of models across countries in the presence of missing forecasts. In the first step, score ratios are computed for all pairs of models by taking the set of overlapping forecasts between the two models and dividing the score of one model by the score achieved by the other model. The relative skill for a given model compared to others is then obtained by taking the geometric mean of all score ratios which involve that model. Low values are better, and the "average" model receives a relative skill score of 1.

Illustration and qualitative observations

When comparing examples of forecasts on the natural scale with those on the log scale (see Fig 4, S7 and S8 Figs) a few interesting patterns emerge. Missing the peak, i.e. predicting increasing numbers while actual observations are already falling, tends to contribute a lot to overall scores on the natural scale (see forecasts during the peak in May 2022 in Fig 4A and 4B). On the log scale, these have less of an influence, as errors are smaller in relative terms (see Fig 4C and 4D). Conversely, failure to predict an upswing while numbers are still low, is less severely penalised on the natural scale (see forecasts in July 2021 and to a lesser extent in July 2022 in Fig 4A and 4B), as overall absolute errors are low. On the log scale, missing lower inflection

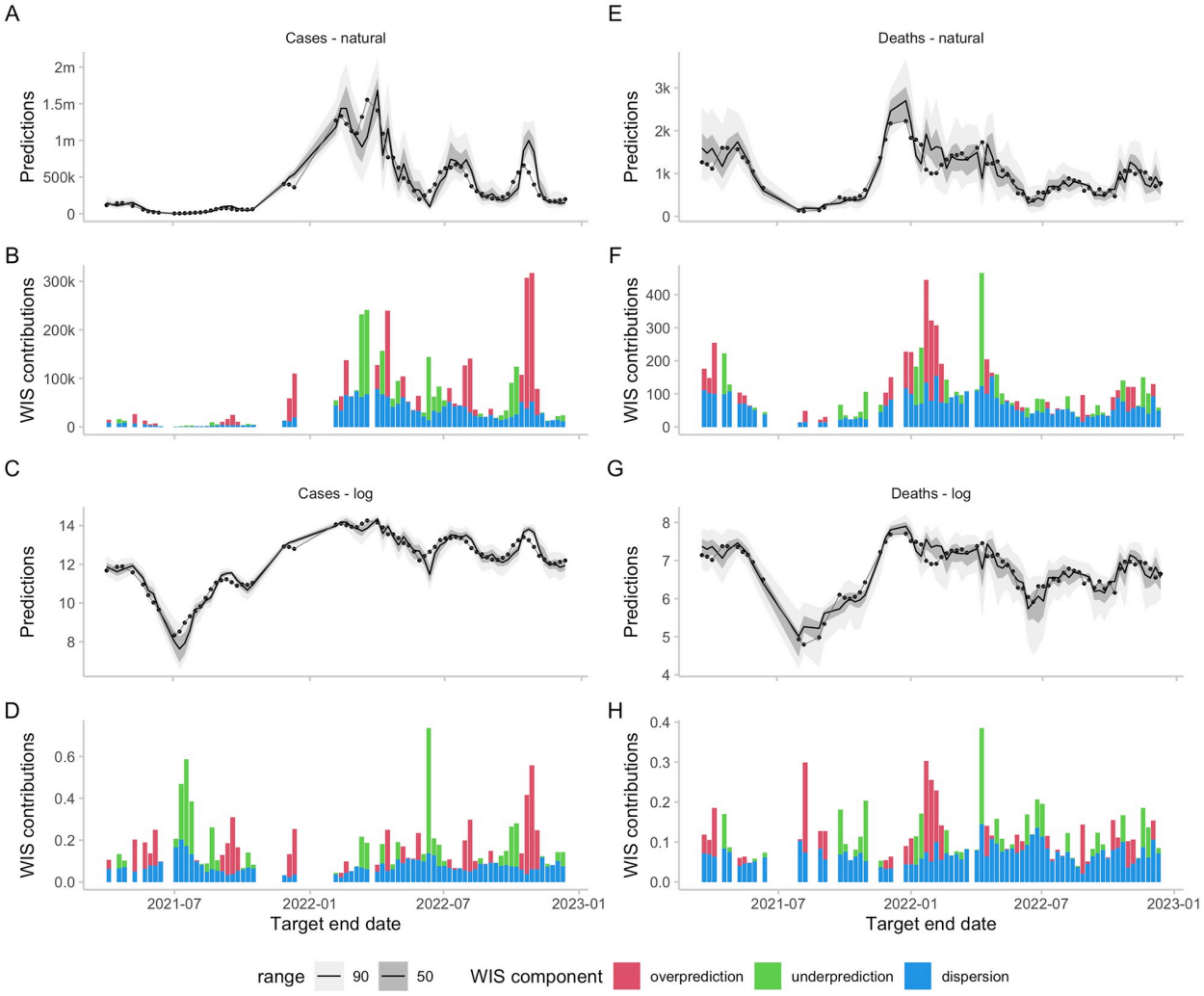

Fig 4. Forecasts and scores for two-week-ahead predictions from the EuroCOVIDhub-ensemble made in Germany. Missing values are due to data anomalies that were removed. A, E: 50% and 90% prediction intervals and observed values for cases and deaths on the natural scale. B, F: Corresponding scores. C, G: Forecasts and observations on the log scale. D, H: Corresponding scores.

https://doi.org/10.1371/journal.pcbi.1011393.g004

points tends to lead to more severe penalties (see Fig 4C and 4D). One can also observe that on the natural scale, scores tend to track the overall level of the target quantity (compare for example forecasts for March-July with forecasts for September-October in Fig 4E and 4F). On the log scale, scores do not exhibit this behaviour and rather increase whenever forecasts are far away from the truth in relative terms, regardless of the overall level of observations.

Across the dataset, the average number of observed cases and deaths varied considerably by location and target type (see Fig 5A and 5B). On the natural scale, scores show a pattern quite similar to the observations across targets (see Fig 5D) and locations (see Fig 5C). On the log scale, scores were more evenly distributed between targets (see Fig 5D) and locations (see Fig 5C). Both on the natural scale as well on the log scale, scores increased considerably with increasing forecast horizon (see Fig 5E). This reflects the increasing difficulty of forecasts further into the future and, for the log scale, corresponds with our expectations based on the theoretical considerations detailed above.

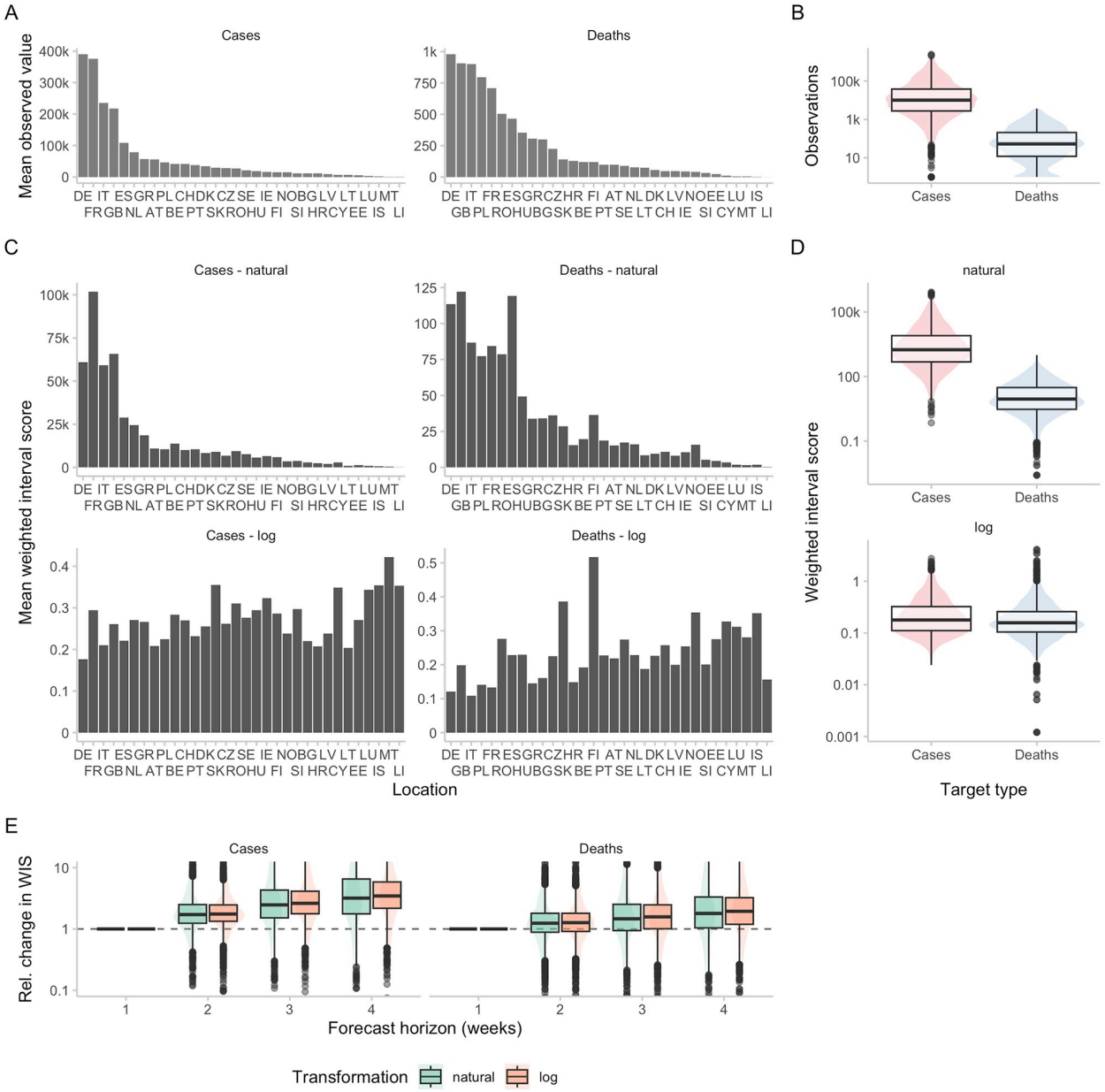

Fig 5. Observations and scores across locations and forecast horizons for the European COVID-19 Forecast Hub data. Locations are sorted according to the mean observed value in that location. A: Average (across all time points) of observed cases and deaths for different locations. B: Corresponding boxplot (y-axis on log-scale) of all cases and deaths. C: Scores for two-week-ahead forecasts from the EuroCOVIDhub-ensemble (averaged across all forecast dates) for different locations, evaluated on the natural scale as well as after transforming counts by adding one and applying the natural logarithm. D: Corresponding boxplots of all individual scores of the EuroCOVIDhub-ensemble for two-week-ahead predictions. E: Boxplots for the relative change of scores for the EuroCOVIDhub-ensemble across forecast horizons. For any given forecast date and location, forecasts were made for four different forecast horizons, resulting in four scores. All scores were divided by the score for forecast horizon one. To enhance interpretability, the range of visible relative changes in scores (relative to horizon = 1) was restricted to [0.1, 10].

https://doi.org/10.1371/journal.pcbi.1011393.g005

To assess the impact of the choice of offset value a we extend the display from Fig 5C by results obtained under different specifications. Results are shown in Fig 6, where for completeness we also added the square root transformation. Smaller values of a increase the relative weight of smaller locations in the overall evaluation. In the most extreme considered

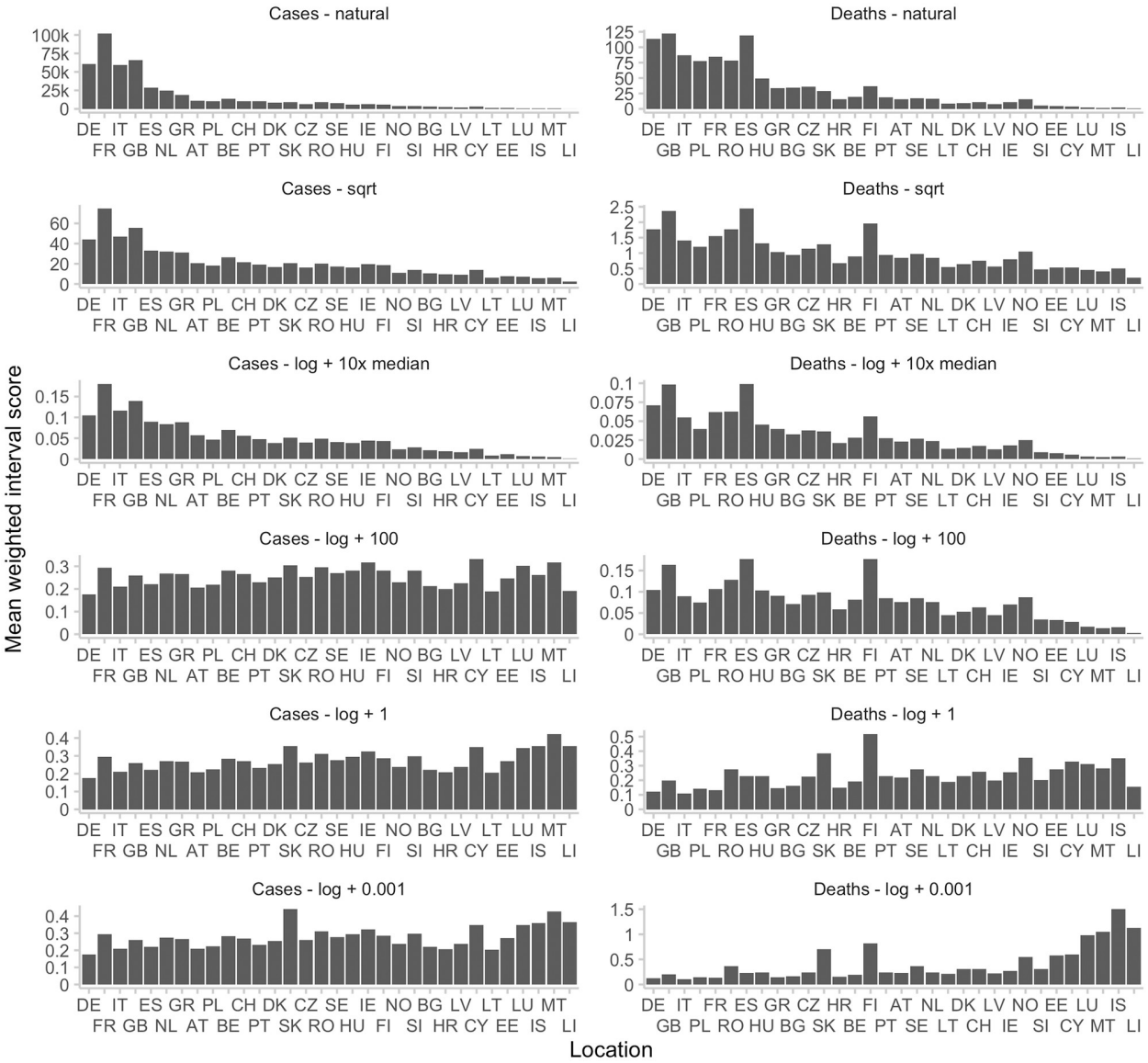

Fig 6. Mean WIS in different locations for different transformations applied before scoring. Locations are sorted according to the mean observed value in that location. Shown are scores for two-week-ahead forecasts of the EuroCOVIDhub-ensemble. On the natural scale (with no transformation prior to applying the WIS), scores correlate strongly with the average number of observed values in a given location. The same is true for scores obtained after applying a square-root transformation, or after applying a log-transformation with a large offset a. For illustrative purposes, a was chosen to be 101630 for cases and 530 for deaths, 10 times the respective median observed value. For large values of a, $\log(x + a)$ grows roughly linearly in x, meaning that we expect to observe the same patterns as in the case with no transformation. For decreasing values of a, we give more relative weight to scores in small locations.

https://doi.org/10.1371/journal.pcbi.1011393.g006

case $a = 0.001$, the smallest locations in fact receive the largest weight both for deaths and cases. For very large values (see the third row of Fig 6), the relative weights strongly resemble those of the evaluation on the natural scale. We recommend using displays of this type to get an intuition for the role different locations may play for overall evaluation results.

Regression analysis to determine the variance-stabilizing transformation

As argued above, the mean-variance, or mean-CRPS, relationship determines which transformation can serve as a VST. We can analyse this relationship empirically by running a regression that explains the WIS (which approximates the CRPS) as a function of the central estimate of the predictive distribution. We ran the regression

$$\log[\mathrm{WIS}(F, y)] = \alpha + \beta \times \log[\mathrm{median}(F)], \tag{18}$$

where the predictive distribution F and the observation y are on the natural scale. This is equivalent to

$$\mathrm{WIS}(F, y) = \exp(\alpha) \times \mathrm{median}(F)^{\beta}, \tag{19}$$

meaning that we estimate a polynomial relationship between the predictive median and achieved WIS. Note that we are using predictive medians rather than means as only the former are available in the European COVID-19 Forecast Hub. As (under the simplifying assumption of normality; see the previous theoretical discussion on the mean-variance relationship) the WIS/CRPS of an ideal forecaster scales with the standard deviation, a value of $\beta = 1$ would imply a quadratic median-variance relationship; the natural logarithm could then serve as a VST. A value of $\beta = 0.5$ would imply a linear median-variance relationship, suggesting the square root as a VST. We applied the regression to case and death forecasts, stratified for one through four-week-ahead forecasts. Results are provided in Table 1. It can be seen that the estimates of β always take a value somewhat below 1, implying a slightly sub-quadratic mean-variance relationship. The logarithmic transformation should thus approximately stabilize the variance (and WIS), possibly leading to somewhat higher scores for smaller forecast targets. The square-root transformation, on the other hand, can be expected to still lead to higher WIS values for targets of higher orders of magnitude.

To check the relationship after the transformation, we ran the regressions

$$\mathrm{WIS}(F_{\log}, \log y) = \alpha_{\log} + \beta_{\log} \cdot \log\left(\mathrm{median}(F)\right), \tag{20}$$

where $F_{\log}$ is the predictive distribution for $\log(y)$, and

$$\mathrm{WIS}(F_{\sqrt{\ }}, \sqrt{y}) = \alpha_{\sqrt{\ }} + \beta_{\sqrt{\ }} \cdot \sqrt{\mathrm{median}(F)}, \tag{21}$$

where $F_{\sqrt{\ }}$ is the predictive distribution on the square-root scale. A value of $\beta_{\log} = 0$ (or $\beta_{\sqrt{\ }} = 0$, respectively) would imply that scores are linearly independent of the median

Table 1. Coefficients of three regressions for the effect of the magnitude of the median forecast on expected scores. The first regression was $\log[\mathrm{WIS}(F, y)] = \alpha + \beta \times \log[\mathrm{median}(F)]$, where F is the predictive distribution and y the observed value.The second one was $\mathrm{WIS}(F_{\log}, \log y) = \alpha_{\log} + \beta_{\log} \cdot \log(\mathrm{median}(F))$, where $F_{\log}$ is the predictive distribution for $\log y$. The third one was $\mathrm{WIS}(F_{\sqrt{\ }}, \sqrt{y}) = \alpha_{\sqrt{\ }} + \beta_{\sqrt{\ }} \cdot \sqrt{(\mathrm{median}(F))}$, where $F_{\sqrt{\ }}$ is the predictive distribution for $\sqrt{y}$.

Horizon	Target	α	β	$\alpha_{\sqrt{\ }}$	$\beta_{\sqrt{\ }}$	$\alpha_{\log}$	$\beta_{\log}$
1	Cases	-0.862	0.876	0.790	0.087	0.433	-0.024
2	Cases	-0.243	0.877	0.959	0.162	0.660	-0.031
3	Cases	0.372	0.855	1.109	0.238	0.882	-0.037
4	Cases	0.816	0.837	1.645	0.296	1.009	-0.036
1	Deaths	-1.146	0.832	0.457	0.048	0.376	-0.035
2	Deaths	-0.981	0.867	0.443	0.084	0.416	-0.028
3	Deaths	-0.807	0.885	0.349	0.131	0.453	-0.019
4	Deaths	-0.602	0.891	0.125	0.194	0.501	-0.011

https://doi.org/10.1371/journal.pcbi.1011393.t001

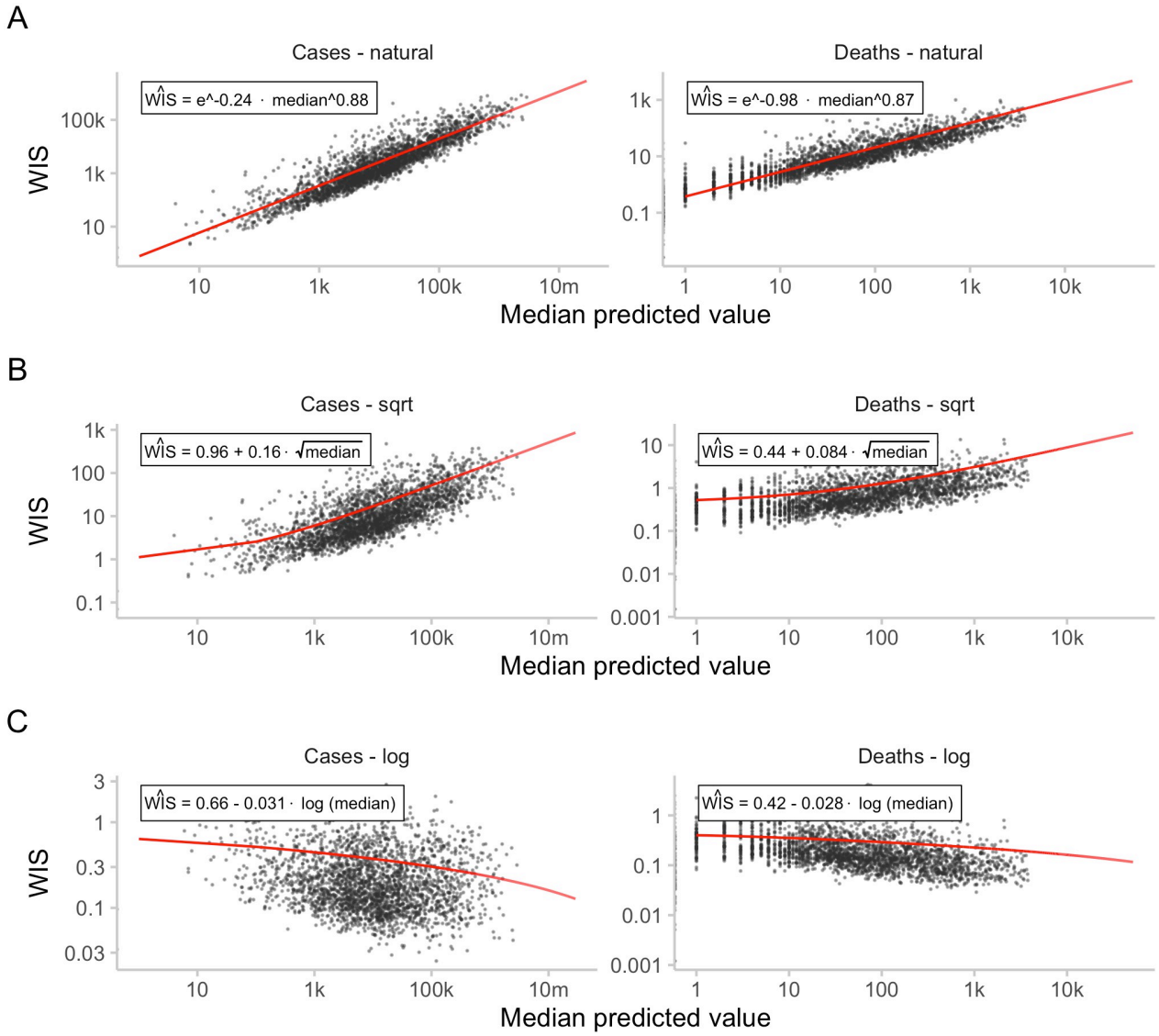

Fig 7. Relationship between median forecasts and scores. Black dots represent WIS values for two-week ahead predictions of the EuroCOVIDhub-ensemble. Drawn in red are the regression lines as discussed in the main text and shown in Table 1. A: WIS for two-week-ahead predictions of the EuroCOVIDhub-ensemble against median predicted values. B: Same as A, with scores obtained after applying a square-root-transformation to the data. C: Same as A, with scores obtained after applying a log-transformation to the data.

https://doi.org/10.1371/journal.pcbi.1011393.g007

prediction after the transformation. A value smaller (larger) than 0 would imply that smaller (larger) targets lead to higher scores. As can be seen from Table 1, the results indeed indicate that small targets lead to larger average WIS when using the log transform ($\beta_{\log} < 0$), while the opposite is true for the square-root transform ($\beta_{\sqrt{}} > 0$). The results of the three regressions are also displayed in Fig 7. In this empirical example, the log transformation thus helps (albeit not perfectly), to stabilise WIS values, and it does so more successfully than the square-root transformation. As can be seen from Fig 7, the expected WIS scores for case targets with medians of 10 and 100,000 differ by more then a factor of ten for the square root transformation, but only a factor of around 2 for the logarithm.

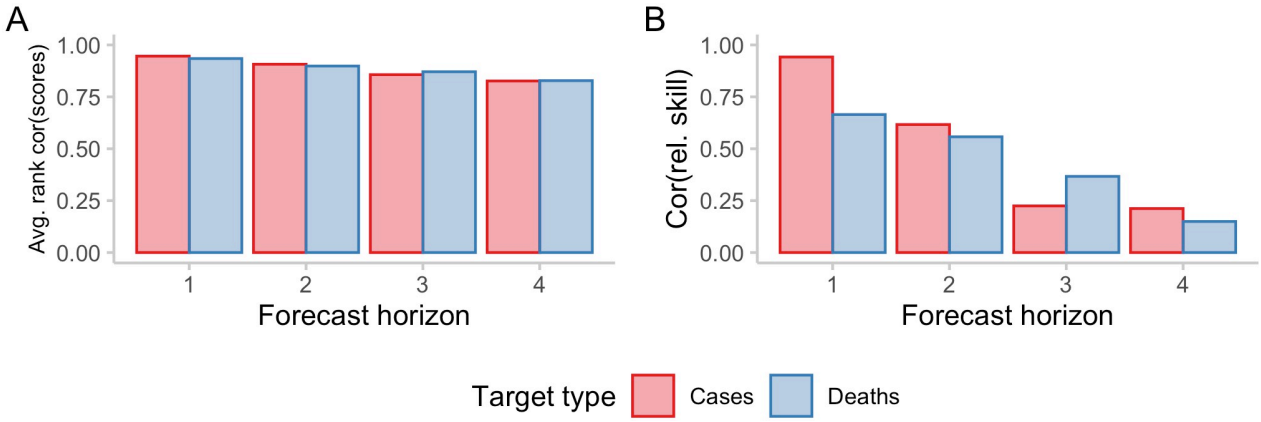

Fig 8. Correlations of rankings on the natural and logarithmic scale. A: Average Spearman rank correlation of scores for individual forecasts. For every individual target (defined by a combination of forecast date, target type, horizon, location), one score was obtained per model. Then, for every forecast target, the Spearman rank correlation was computed between scores on the natural scale and on the log scale for all the models that had made a forecast for that specific target. These individual rank correlations were then averaged across locations and time and are displayed stratified by horizon and target types, representing average accordance of model ranks for a single forecast target on the natural and on the log scale. B: Correlation between relative skill scores. For every forecast horizon and target type, a separate relative skill score was computed per model using pairwise comparisons, which is a measure of performance of a model relative to the others for a given horizon and target type that accounts for missing values. The plot shows the correlation between the relative skill scores on the natural vs. on the log scale, representing accordance of overall model performance as judged by scores on the natural and on the log scale.

https://doi.org/10.1371/journal.pcbi.1011393.g008

Impact of logarithmic transformation on model rankings

For *individual* forecasts, rankings between models for single forecasts are mostly preserved, with differences increasing across forecast horizons (see Fig 8A). While rankings between forecasters remain similar for a single forecast, this is not true anymore when looking at rankings obtained after averaging scores across multiple forecasts made at different times or in different locations. As discussed earlier, scores on the natural and on the log scale penalise errors very differently, e.g. when looking at performance during peaks or troughs. When evaluating performance *averaged across* different forecasts and forecast targets, relative skill scores of the models therefore change considerably (Fig 8B). The correlation between relative skill scores also decreases noticeably with increasing forecast horizon.

Fig 9 shows the changes in the ranking between different forecasting models. Encouragingly for the European Forecast Hub, the Hub ensemble, which is the forecast the organisers suggest forecast consumers make use of, remains the top model across scoring schemes. For cases, the ILM-EKF model and the Forecast Hub baseline model exhibit the largest change in relative skill scores. For the ILM-EKF model the relative proportion of the score that is due to overprediction is reduced when applying a log-transformation before scoring (see Fig 9E. Instances where the model has overshot are penalised less heavily on the log scale, leading to an overall better score. For the Forecast Hub baseline model, the fact that it often puts relevant probability mass on zero (see S7 Fig), leads to worse scores after applying log-transformation due to large dispersion penalties. For deaths, the baseline model seems to get similarly penalised for its in relative terms highly dispersed forecasts. The performance of other models changes as well, but patterns are less discernible on this aggregate level.

Discussion

In this paper, we proposed the use of transformations, with a particular focus on the natural logarithmic transformation, when evaluating forecasts in an epidemiological setting. These

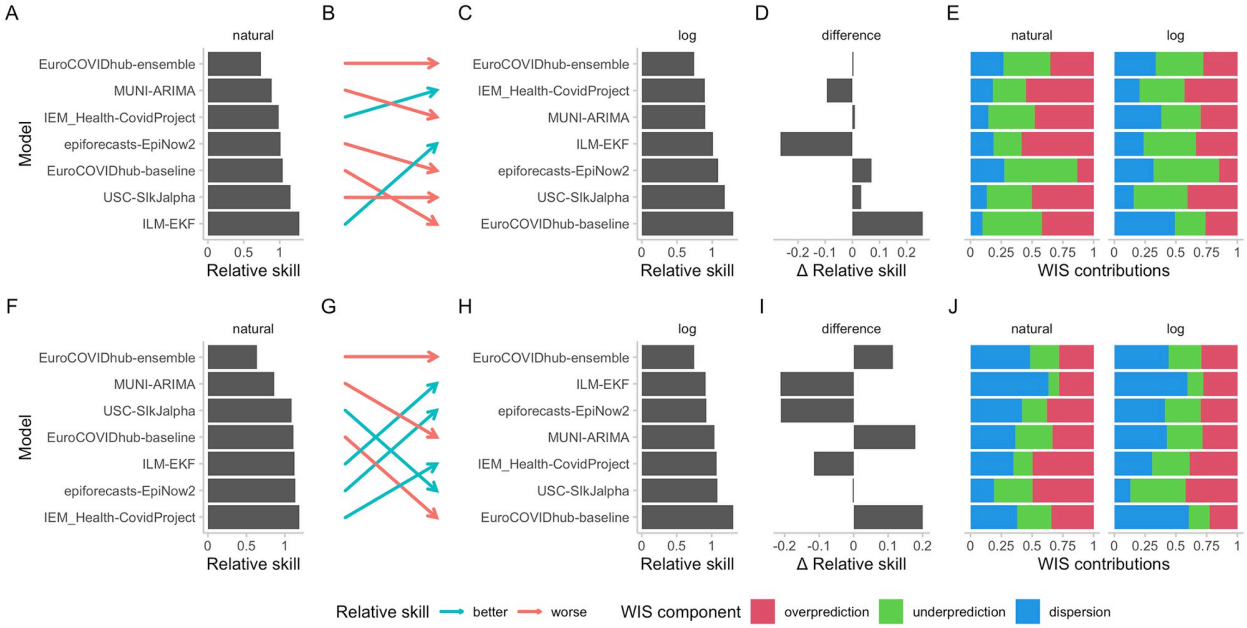

Fig 9. Changes in model ratings as measured by relative skill for two-week-ahead predictions for cases (top row) and deaths (bottom row). A: Relative skill scores for case forecasts from different models submitted to the European COVID-19 Forecast Hub computed on the natural scale. B: Change in rankings as determined by relative skill scores when moving from an evaluation on the natural scale to one on the logarithmic scale. Red arrows indicate that the relative skill scores deteriorated when moving from the natural to the log scale, green arrows indicate they improved. C: Relative skill scores based on scores on the log scale. D: Difference in relative skill scores computed on the natural and on the logarithmic scale, ordered as in C. E: Relative contributions of the different WIS components (overprediction, underprediction, and dispersion) to overall model scores on the natural and the logarithmic scale. F, G, H, I, J: Analogously for deaths.

https://doi.org/10.1371/journal.pcbi.1011393.g009

transformations can address issues that arise when evaluating epidemiological forecasts based on measures of absolute error and their probabilistic generalisations (i.e CRPS and WIS). We showed that scores obtained after log-transforming both forecasts and observations can be interpreted as a) a measure of relative prediction errors, as well as b) a score for a forecast of the exponential growth rate of the target quantity and c) as variance stabilising transformation in some settings. When applying this approach to forecasts from the European COVID-19 Forecast Hub, we found overall scores on the log scale to be more equal across, time, location and target type (cases, deaths) than scores on the natural scale. Scores on the log scale were much less influenced by the overall incidence level in a country and showed a slight tendency to be higher in locations with very low incidences. We found that model rankings changed noticeably.

On the natural scale, missing the peak and overshooting was more severely penalised than missing the nadir and the following upswing in numbers. Both failure modes tended to be more equally penalised on the log scale (with undershooting receiving slightly higher penalties in our example).

Applying a log-transformation prior to the WIS means that forecasts are evaluated in terms of relative errors and errors on the exponential growth rate, rather than absolute errors. The most important strength of this approach is that the evaluation better accommodates the exponential nature of the epidemiological process and the types of errors forecasters who accurately model those processes are expected to make. The log-transformation also helps avoid issues with scores being strongly influenced by the order of magnitude of the forecast quantity, which can be an issue when evaluating forecasts on the natural scale. A potential downside is that forecast evaluation is unreliable in situations where observed values are zero or very small. One

could argue that this correctly reflect inherent uncertainty about the future course of an epidemic when numbers are small. Users nevertheless need to be aware that this can pose issues in practice. Including very small values in prediction intervals (see S7 Fig for an example) can lead to excessive dispersion values on the log scale. Similarly, locations with lower incidences may get disproportionate weight (i.e. high scores) when evaluating forecasts on the log scale. [8] argue that it is desirable to give large weight to forecasts for locations with high incidences, as this reflects performance on the targets we should care about most. On the other hand, scoring forecasts on the log scale may be less influenced by outliers and better reflect consistent performance across time, space, and forecast targets. Furthermore, decision makers may specifically care about situations in which numbers start to rise from a previously low level.

The log-transformation is only one of many transformations that may be useful and appropriate in an epidemiological context. One obvious option is to apply a population standardization to obtain incidence forecasts e.g., per 100,000 population [35]. We suggested using the natural logarithm as a variance-stabilising transformation (VST). This is appropriate for variables that are approximately normally distributed and have a quadratic mean-variance relationship with $\sigma^2 = c \times \mu^2$ (this is e.g. approximately true for the negative binomoial distribution and large μ). Alternatively, the square-root transformation can be appropriate in the case of a Poisson distributed variable [30]. Other VST like the Box-Cox [36] are conceivable as well. If one is interested in multiplicative, rather than exponential growth rates, one could, instead of applying a log transformation, convert forecasts into forecasts for the multiplicative growth rate by dividing numbers by the last value that was observed at the time the forecast was made. Forecasters would then implicitly predict a separate multiplicative growth rate from today to horizon 1, 2, etc. Instead of dividing by the last observed value, another promising transformation would be to divide each forecast by the forecast of the previous week (and analogously for observations), in order to obtain forecasts for week-to-week growth rates. Alternatively, one could also take first differences of values on the log scale. This approach would be akin to evaluating the shape of the predicted trajectory against the shape of the observed trajectory (for a different approach to evaluating the shape of a forecast, see [37]). Dividing values by the previous value, unfortunately, is not feasible under the current quantile-based format of the Forecast Hubs, as the growth rate of the α-quantile may be different from the α-quantile of the growth-rate. However, it may be an interesting approach if predictive samples are available or if quantiles for weekwise growth rates have been collected. Potentially, the variance stabilising time-series forecasting literature may be a useful source of other transformations for various forecast settings.

It is possible to go beyond choosing a single transformation by constructing composite scores as a weighted sum of scores based on different transformations. This would make it possible to create custom scores and allow forecast consumers to choose and assign explicit weights to different qualities of the forecasts they might care about.

Exploring transformations is a promising avenue for future work that could help bridge the gap between modellers and policymakers by providing scoring rules that better reflect what forecast consumers care about. In this paper, we did not make any particular assumptions about policy makers' priorities and preferences. Rather, we aimed to enable users to make an informed choice by showing how different transformations lead to different relative weights for the kinds of prediction errors forecast consumers may care about, such as absolute vs. relative errors or the size of penalties for over- vs. underprediction. In practice, engagement with decision makers is important to determine what their priorities are and how different ways to measure predictive importance should be weighed.

We have shown that the natural logarithm transformation can lead to significant changes in the relative rankings of models against each other, with potentially important implications for decision-makers who rely on the knowledge of past performance to make a judgement about

which forecasts should inform future decisions. While it is commonly accepted that multiple proper scoring rules should usually be considered when comparing forecasts, we think this should be supplemented by considering different transformations of the data to obtain a richer picture of model performance. More work needs to be done to better understand the effects of applying transformations in different contexts, and how they may impact decision-making.

Supporting information

S1 Text. Alternative Formulation of the WIS.
(PDF)

S1 Table. Summary statistics for observations and scores for forecasts from the ECDC data set.
(PDF)

S2 Table. Criteria for removing forecasts. Any forecast that met one of the listed criteria (represented by a row in the table), was removed. Those forecasts were removed in order to be better able to illustrate the effects of the log-transformation on scores and eliminating distortions caused by outlier forecasters. When evaluating models against each other (rather than illustrating the effect of a transformation), one would prefer not to condition on the outcome when deciding whether a forecast should be taken into account.
(PDF)

S1 Fig. Illustration of the effect of applying a transformation after scoring. We assume $Y \sim$ LogNormal(0, 1) and evaluate the expected CRPS for predictive distributions LogNormal(0, σ) with varying values of $\sigma \in [0.1, 2]$. For the regular CRPS (left) and CRPS applied to log-transformed outcomes (middle), the lowest expectation is achieved for the true value $\sigma = 1$. For the log-transformed CRPS, the optimal value is 0.9, i.e. there is an incentive to report a forecast that is too sharp. The score is therefore no longer proper.
(TIF)

S2 Fig. Illustration of the effect of adding a small quantity to a value before taking the natural logarithm. For increasing x, all lines eventually approach the black line (representing a transformation with no offset applied). For a given solid line, the dashed line of the same colour marks the x-value that is equal to 5 times the corresponding offset. It can be seen that for a values smaller than one fifth of the transformed quantity, the effect of adding an offset is generally small. When choosing a suitable a, the trade-off is between staying close to the interpretation of a pure log-transformation (choosing a small a) and not giving excessive weights to small observations (by choosing a larger a, see Fig 6).
(TIF)

S3 Fig. Visualisation of expected CRPS values against approximated scores. This is using the approximation detailed in theoretical discussion on model rankings (see also Fig 2). Expected CRPS scores are shown for three different distributions once on the natural scale (top row) and once scored on the log scale (bottom row).
(TIF)

S4 Fig. Number of forecasts available from different models for each forecast date.
(TIF)

S5 Fig. Number of observed values that were removed as anomalous. The values were marked as anomalous by the European Forecast Hub team.
(TIF)

S6 Fig. Number of forecasts marked as erroneous and removed. Forecasts that were in extremely poor agreement with the observed values were removed from the analysis according to the criteria shown in S2 Table.
(TIF)

S7 Fig. Forecasts and scores for two-week-ahead predictions from the EuroCOVIDhub-baseline made in Germany. The model had zero included in some of its 50 percent intervals (e.g. for case forecasts in July 2021), leading to excessive dispersion values on the log scale. One could argue that including zero in the prediction intervals constituted an unreasonable forecast that was rightly penalised, but in general care has to be taken with small numbers. One potential way to do deal with this could be to use a higher a value when applying a transformation $\log(x + a)$, for example $a = 10$ instead of $a = 1$. A, E: 50% and 90% prediction intervals and observed values for cases and deaths on the natural scale. B, F: Corresponding scores. C, G: Forecasts and observations on the log scale. D, H: Corresponding scores.
(TIF)

S8 Fig. Forecasts and scores for two-week-ahead predictions from the epiforecasts-EpiNow2 model made in Germany. A, E: 50% and 90% prediction intervals and observed values for cases and deaths on the natural scale from the EpiNow2 model [38]. B, F: Corresponding scores. C, G: Forecasts and observations on the log scale. D, H: Corresponding scores.
(TIF)

Author Contributions

Conceptualization: Nikos I. Bosse, Sam Abbott, Anne Cori, Edwin van Leeuwen, Johannes Bracher, Sebastian Funk.

Data curation: Nikos I. Bosse.

Formal analysis: Nikos I. Bosse, Sam Abbott, Johannes Bracher, Sebastian Funk.

Investigation: Nikos I. Bosse, Sam Abbott, Johannes Bracher, Sebastian Funk.

Methodology: Nikos I. Bosse, Sam Abbott, Johannes Bracher, Sebastian Funk.

Project administration: Nikos I. Bosse.

Resources: Sebastian Funk.

Software: Nikos I. Bosse.

Supervision: Sam Abbott, Anne Cori, Edwin van Leeuwen, Johannes Bracher, Sebastian Funk.

Validation: Nikos I. Bosse, Sam Abbott, Johannes Bracher.

Visualization: Nikos I. Bosse, Johannes Bracher.

Writing – original draft: Nikos I. Bosse.

Writing – review & editing: Nikos I. Bosse, Sam Abbott, Anne Cori, Edwin van Leeuwen, Johannes Bracher, Sebastian Funk.

References

1. Held L, Meyer S, Bracher J. Probabilistic Forecasting in Infectious Disease Epidemiology: The 13th Armitage Lecture. Statistics in Medicine. 2017; 36(22):3443–3460. https://doi.org/10.1002/sim.7363 PMID: 28656694

2. Reich NG, Lessler J, Funk S, Viboud C, Vespignani A, Tibshirani RJ, et al. Collaborative Hubs: Making the Most of Predictive Epidemic Modeling. American Journal of Public Health. 2022; 112(6):839–842. https://doi.org/10.2105/AJPH.2022.306831 PMID: 35420897
3. Timmermann A. Forecasting Methods in Finance. Annual Review of Financial Economics. 2018; 10 (1):449–479. https://doi.org/10.1146/annurev-financial-110217-022713
4. Gneiting T, Raftery AE. Weather Forecasting with Ensemble Methods. Science. 2005; 310(5746):248–249.
5. Gneiting T, Raftery AE. Strictly Proper Scoring Rules, Prediction, and Estimation. Journal of the American Statistical Association. 2007; 102(477):359–378. https://doi.org/10.1198/016214506000001437
6. Good IJ. Rational Decisions. Journal of the Royal Statistical Society Series B (Methodological). 1952; 14(1):107–114. https://doi.org/10.1111/j.2517-6161.1952.tb00104.x
7. Winkler RL, Muñoz J, Cervera JL, Bernardo JM, Blattenberger G, Kadane JB, et al. Scoring Rules and the Evaluation of Probabilities. Test. 1996; 5(1):1–60. https://doi.org/10.1007/BF02562681
8. Bracher J, Ray EL, Gneiting T, Reich NG. Evaluating Epidemic Forecasts in an Interval Format. PLoS computational biology. 2021; 17(2):e1008618. https://doi.org/10.1371/journal.pcbi.1008618 PMID: 33577550
9. Johansson MA, Apfeldorf KM, Dobson S, Devita J, Buczak AL, Baugher B, et al. An Open Challenge to Advance Probabilistic Forecasting for Dengue Epidemics. Proceedings of the National Academy of Sciences. 2019; 116(48):24268–24274. https://doi.org/10.1073/pnas.1909865116 PMID: 31712420
10. Cramer E, Reich NG, Wang SY, Niemi J, Hannan A, House K, et al. COVID-19 Forecast Hub: 4 December 2020 Snapshot; 2020.
11. Cramer E, Ray EL, Lopez VK, Bracher J, Brennen A, Rivadeneira AJC, et al. Evaluation of Individual and Ensemble Probabilistic Forecasts of COVID-19 Mortality in the US. medRxiv. 2021; p. 2021.02.03.21250974.
12. Sherratt K, Gruson H, Grah R, Johnson H, Niehus R, Prasse B, et al. Predictive Performance of Multi-Model Ensemble Forecasts of COVID-19 across European Nation. Europe PMC. 2022.
13. Bracher J, Wolffram D, Deuschel J, Görgen K, Ketterer JL, Ullrich A, et al. Short-Term Forecasting of COVID-19 in Germany and Poland during the Second Wave—a Preregistered Study. medRxiv. 2021; p. 2020.12.24.20248826.
14. Bracher J, Wolffram D, Deuschel J, Goergen K, Ketterer JL, Ullrich A, et al. National and subnational short-term forecasting of COVID-19 in Germany and Poland, early 2021. Communications Medicine. 2022. https://doi.org/10.1038/s43856-022-00191-8 PMID: 36352249
15. CDC. Cdcepi/Flusight-forecast-data; 2022. CDC Epidemic Prediction Initiative.
16. Gostic KM, McGough L, Baskerville E, Abbott S, Joshi K, Tedijanto C, et al. Practical Considerations for Measuring the Effective Reproductive Number, Rt. medRxiv. 2020. https://doi.org/10.1371/journal.pcbi.1008409 PMID: 33301457
17. Dushoff J, Park SW. Speed and Strength of an Epidemic Intervention. Proceedings of the Royal Society B: Biological Sciences. 2021; 288(1947):20201556. https://doi.org/10.1098/rspb.2020.1556 PMID: 33757359
18. Bolin D, Wallin J. Local Scale Invariance and Robustness of Proper Scoring Rules. Statistical Science. 2023; 38(1):140–159. https://doi.org/10.1214/22-STS864
19. Taylor JW. Evaluating Volatility and Interval Forecasts. Journal of Forecasting. 1999; 18(2):111–128. https://doi.org/10.1002/(SICI)1099-131X(199903)18:2%3C111::AID-FOR713%3E3.0.CO;2-C
20. Mayr J, Ulbricht D. Log versus Level in VAR Forecasting: 42 Million Empirical Answers—Expect the Unexpected. Economics Letters. 2015; 126:40–42. https://doi.org/10.1016/j.econlet.2014.11.008
21. Löwe R, Mikkelsen PS, Madsen H. Stochastic Rainfall-Runoff Forecasting: Parameter Estimation, Multi-Step Prediction, and Evaluation of Overflow Risk. Stochastic Environmental Research and Risk Assessment. 2014; 28(3):505–516. https://doi.org/10.1007/s00477-013-0768-0
22. Fuglstad GA, Simpson D, Lindgren F, Rue H. Does Non-Stationary Spatial Data Always Require Non-Stationary Random Fields? Spatial Statistics. 2015; 14:505–531.
23. Lehmann EL. Some Principles of the Theory of Testing Hypotheses. The Annals of Mathematical Statistics. 1950; 21(1):1–26. https://doi.org/10.1214/aoms/1177729884
24. Diks C, Panchenko V, van Dijk D. Likelihood-Based Scoring Rules for Comparing Density Forecasts in Tails. Journal of Econometrics. 2011; 163(2):215–230. https://doi.org/10.1016/j.jeconom.2011.04.001
25. European Covid-19 Forecast Hub. European Covid-19 Forecast Hub; 2021. https://covid19forecasthub.eu/.
26. Gneiting T. Making and Evaluating Point Forecasts. Journal of the American Statistical Association. 2011; 106(494):746–762. https://doi.org/10.1198/jasa.2011.r10138

27. Flores BE. A pragmatic view of accuracy measurement in forecasting. Omega. 1986; 14(2):93–98. https://doi.org/10.1016/0305-0483(86)90013-7
28. Wallinga J, Lipsitch M. How Generation Intervals Shape the Relationship between Growth Rates and Reproductive Numbers. Proceedings of the Royal Society B: Biological Sciences. 2007; 274 (1609):599–604. https://doi.org/10.1098/rspb.2006.3754 PMID: 17476782
29. Bartlett MS. The Square Root Transformation in Analysis of Variance. Supplement to the Journal of the Royal Statistical Society. 1936; 3(1):68–78. https://doi.org/10.2307/2983678
30. Dunn PK, Smyth GK. Generalized Linear Models With Examples in R. Springer; 2018.
31. Bellégo C, Benatia D, Pape L. Dealing with Logs and Zeros in Regression Models; 2022.
32. Lerch S, Thorarinsdottir TL, Ravazzolo F, Gneiting T. Forecaster's Dilemma: Extreme Events and Forecast Evaluation; 2015.
33. R Core Team. R: A Language and Environment for Statistical Computing; 2022. Available from: https://www.R-project.org/.
34. Bosse NI, Gruson H, Cori A, van Leeuwen E, Funk S, Abbott S. Evaluating Forecasts with Scoringutils in R. arXiv. 2022.
35. Abbott S, Sherratt K, Bosse N, Gruson H, Bracher J, Funk S. Evaluating an Epidemiologically Motivated Surrogate Model of a Multi-Model Ensemble; 2022.
36. Box GEP, Cox DR. An Analysis of Transformations. Journal of the Royal Statistical Society Series B (Methodological). 1964; 26(2):211–252. https://doi.org/10.1111/j.2517-6161.1964.tb00553.x
37. Srivastava A, Singh S, Lee F. Shape-Based Evaluation of Epidemic Forecasts; 2022.
38. Abbott S, Hellewell J, Sherratt K, Gostic K, Hickson J, Badr HS, et al. EpiNow2: Estimate Real-Time Case Counts and Time-Varying Epidemiological Parameters; 2020.

A Supplementary information

A.1 Alternative Formulation of the WIS

Instead of defining the WIS as an average of scores for individual quantiles, we can define it using an average of scores for symmetric predictive intervals. For a single prediction interval, the interval score (IS) is computed as the sum of three penalty components, dispersion (width of the prediction interval), underprediction and overprediction,

$$IS_\alpha(F, y) = (u - l) + \frac{2}{\alpha} \cdot (l - y) \cdot \mathbf{1}(y \leq l) + \frac{2}{\alpha} \cdot (y - u) \cdot \mathbf{1}(y \geq u) \tag{1}$$

$$= \text{dispersion} + \text{underprediction} + \text{overprediction}, \tag{2}$$

where $\mathbf{1}()$ is the indicator function, y is the observed value, and l and u are the $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$ quantiles of the predictive distribution, i.e. the lower and upper bound of a single central prediction interval. For a set of K^* prediction intervals and the median m, the WIS is computed as a weighted sum,

$$\text{WIS} = \frac{1}{K^* + 0.5} \cdot \left(w_0 \cdot |y - m| + \sum_{k=1}^{K^*} w_k \cdot IS_{\alpha_k}(F, y) \right), \tag{3}$$

where w_k is a weight for every interval. Usually, $w_k = \frac{\alpha_k}{2}$ and $w_0 = 0.5$.

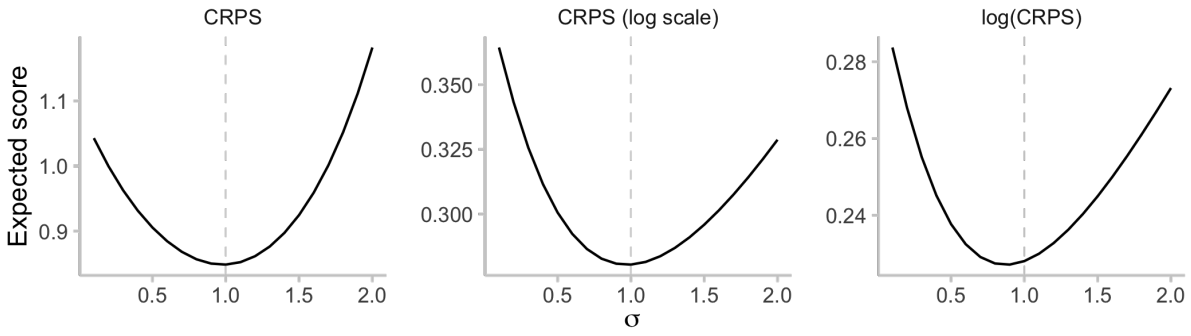

Figure SI.1: Illustration of the effect of applying a transformation after scoring. We assume $Y \sim \text{LogNormal}(0, 1)$ and evaluate the expected CRPS for predictive distributions $\text{LogNormal}(0, \sigma)$ with varying values of $\sigma \in [0.1, 2]$. For the regular CRPS (left) and CRPS applied to log-transformed outcomes (middle), the lowest expectation is achieved for the true value $\sigma = 1$. For the log-transformed CRPS, the optimal value is 0.9, i.e. there is an incentive to report a forecast that is too sharp. The score is therefore no longer proper.

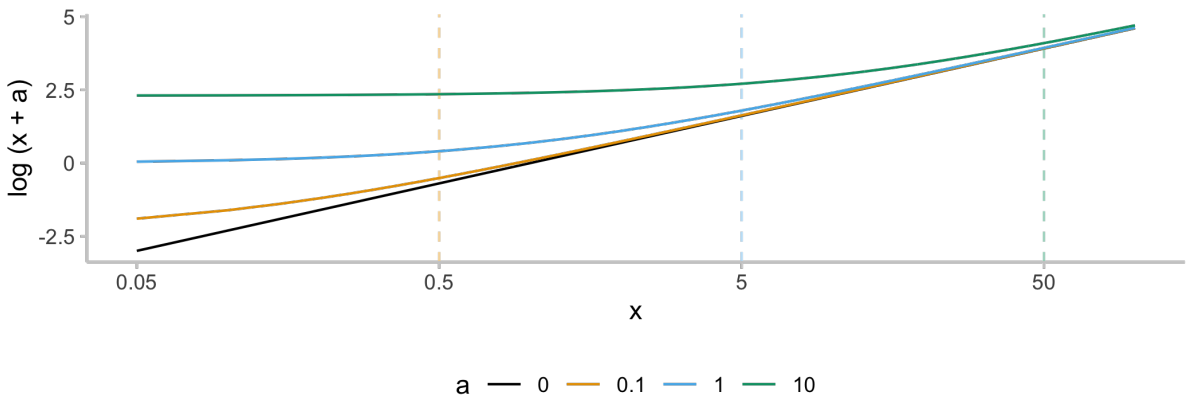

Figure SI.2: Illustration of the effect of adding a small quantity to a value before taking the natural logarithm. For increasing x, all lines eventually approach the black line (representing a transformation with no offset applied). For a given solid line, the dashed line of the same colour marks the x-value that is equal to 5 times the corresponding offset. It can be seen that for a values smaller than one fifth of the transformed quantity, the effect of adding an offset is generally small. When choosing a suitable a, the trade-off is between staying close to the interpretation of a pure log-transformation (choosing a small a) and not giving excessive weights to small observations (by choosing a larger a, see Figure 6).

target type	quantity	measure	natural	log
Cases	Observations	mean	61979	9.19
Cases	Observations	sd	171916	2.10
Cases	Observations	var	29555122130	4.42
Deaths	Observations	mean	220	3.89
Deaths	Observations	sd	435	1.96
Deaths	Observations	var	189051	3.83
Cases	WIS	mean	15840	0.27
Cases	WIS	sd	53117	0.28
Deaths	WIS	mean	31	0.23
Deaths	WIS	sd	65	0.28

Table SI.1: Summary statistics for observations and scores for forecasts from the ECDC data set.

True value	&	Median prediction
> 0		$> 100\times$ true value
> 10		$> 20\times$ true value
> 50		$< 1/50\times$ true value
$= 0$		> 100

Table SI.2: Criteria for removing forecasts. Any forecast that met one of the listed criteria (represented by a row in the table), was removed. Those forecasts were removed in order to be better able to illustrate the effects of the log-transformation on scores and eliminating distortions caused by outlier forecasters. When evaluating models against each other (rather than illustrating the effect of a transformation), one would prefer not to condition on the outcome when deciding whether a forecast should be taken into account.

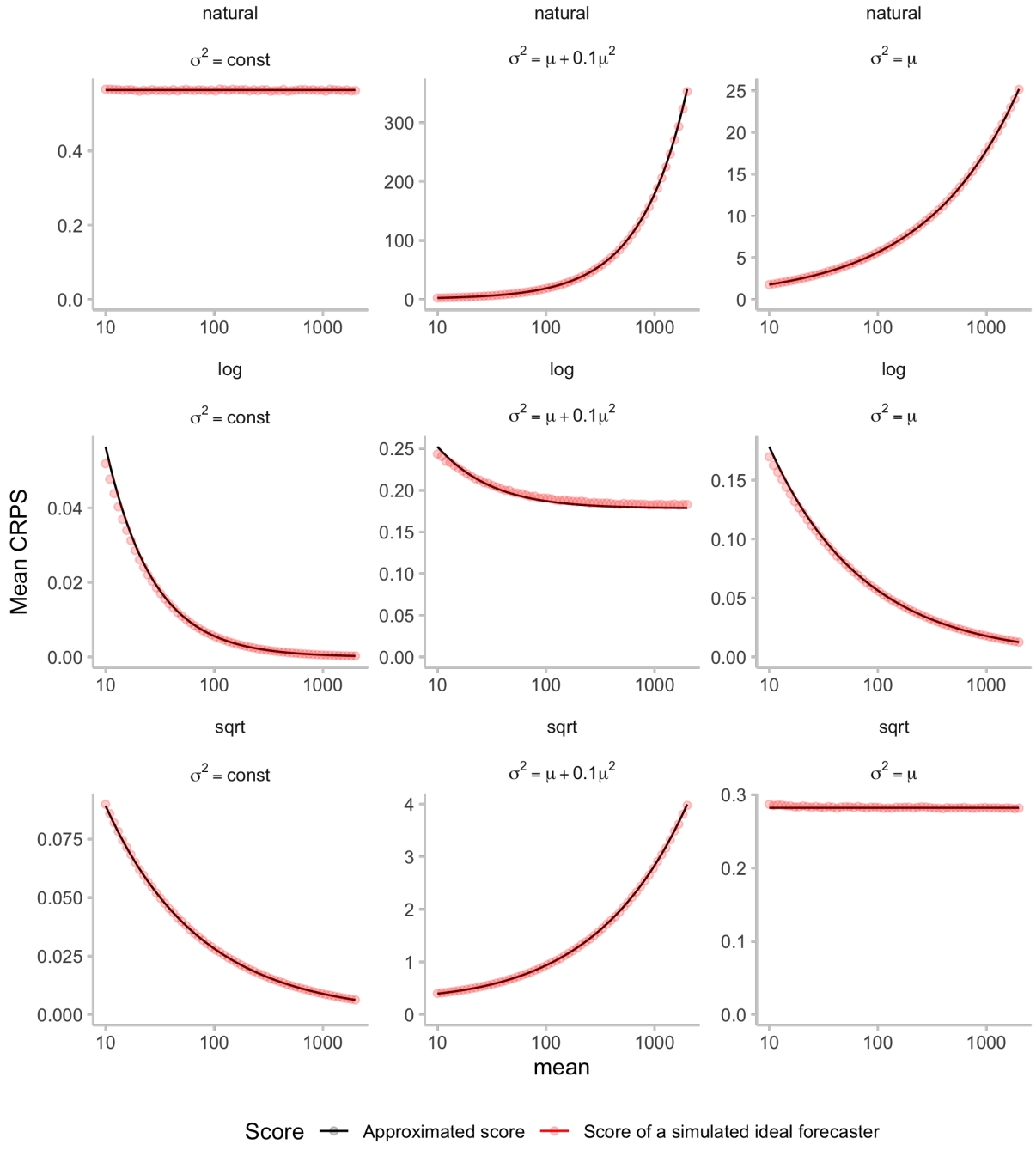

Figure SI.3: This is using the approximation detailed in the theoretical discussion on model rankings (see also Fig 2). Expected CRPS scores are shown for three different distributions once on the natural scale (top row) and once scored on the log scale (bottom row).

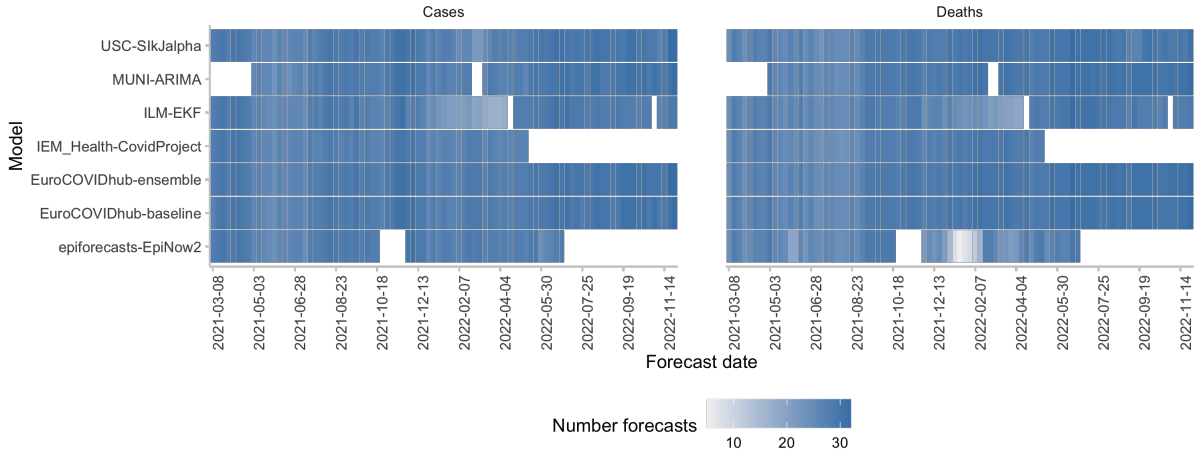

Figure SI.4: Number of forecasts available from different models for each forecast date.

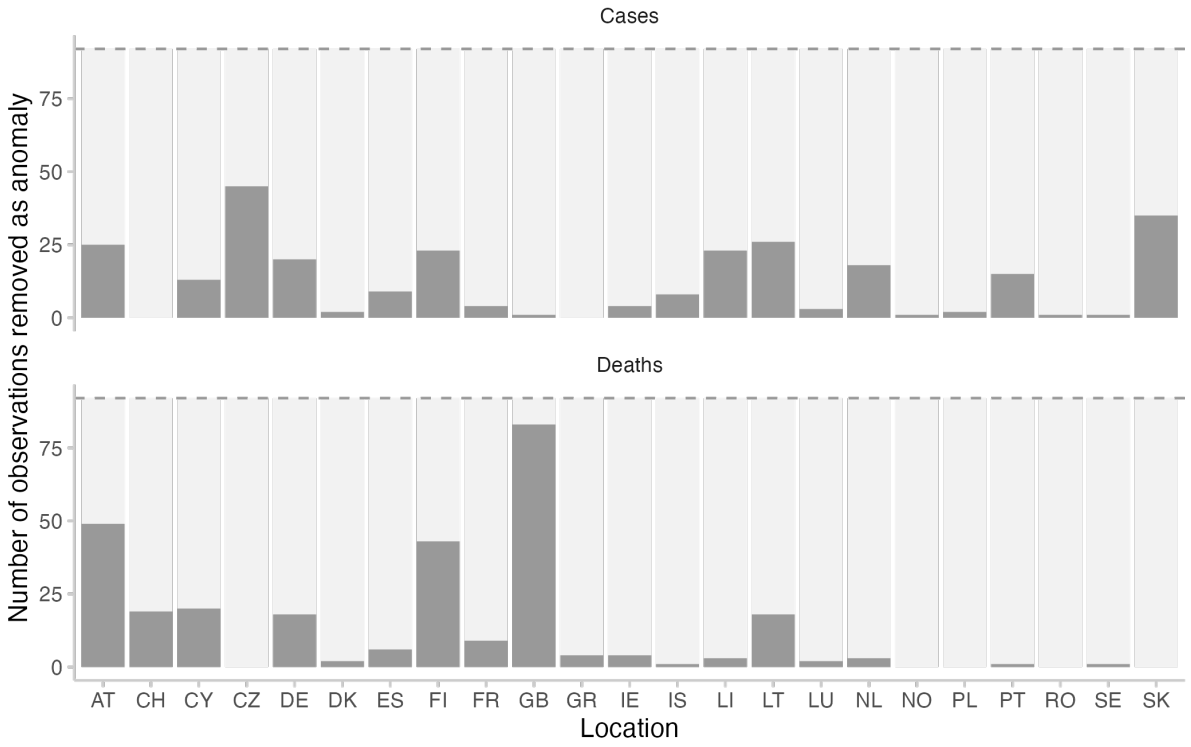

Figure SI.5: Number of observed values that were removed as anomalous. The values were marked as anomalous by the European Forecast Hub team.

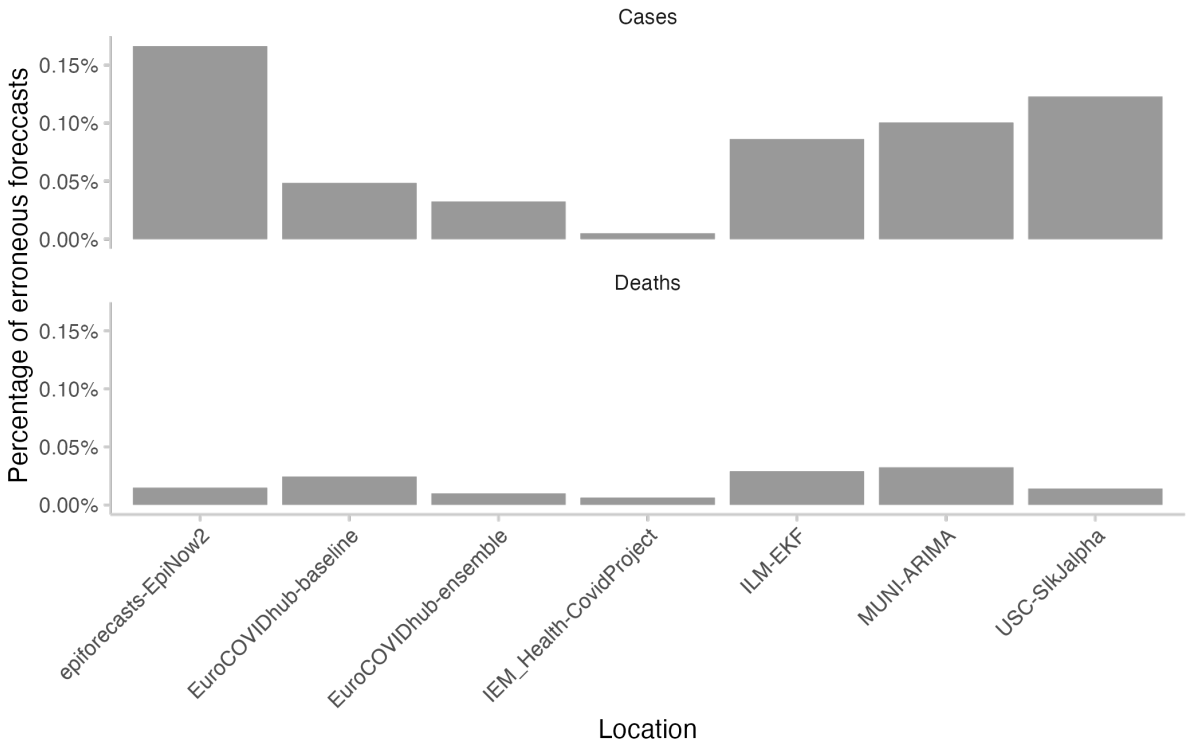

Figure SI.6: Number of forecasts marked as erroneous and removed. Forecasts that were in extremely poor agreement with the observed values were removed from the analysis according to the criteria shown in Table SI.2.

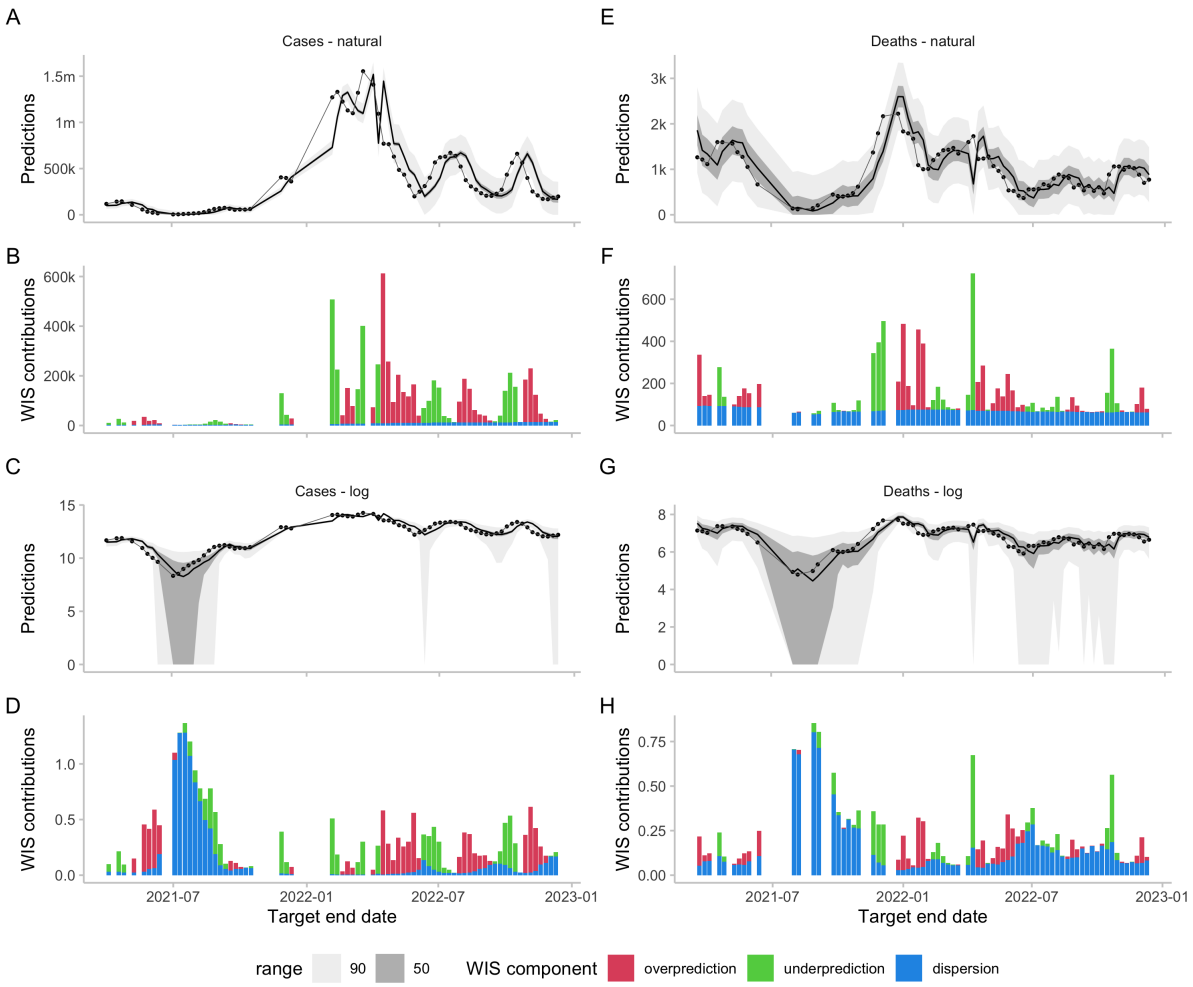

Figure SI.7: Forecasts and scores for two-week-ahead predictions from the EuroCOVIDhub-baseline made in Germany. The model had zero included in some of its 50 percent intervals (e.g. for case forecasts in July 2021), leading to excessive dispersion values on the log scale. One could argue that including zero in the prediction intervals constituted an unreasonable forecast that was rightly penalised, but in general care has to be taken with small numbers. One potential way to do deal with this could be to use a higher a value when applying a transformation $\log(x + a)$, for example $a = 10$ instead of $a = 1$. A, E: 50% and 90% prediction intervals and observed values for cases and deaths on the natural scale. B, F: Corresponding scores. C, G: Forecasts and observations on the log scale. D, H: Corresponding scores.

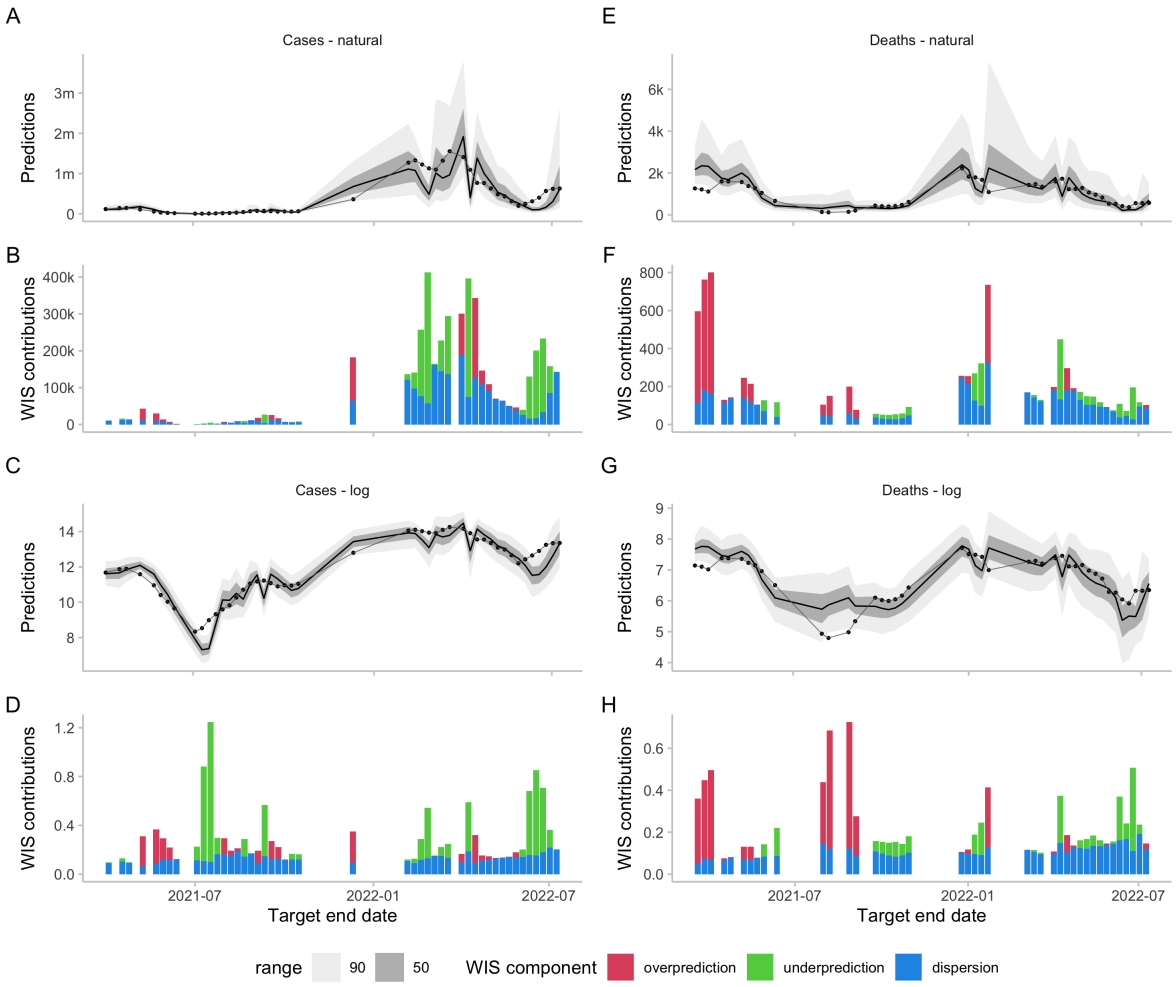

Figure SI.8: Forecasts and scores for two-week-ahead predictions from the epiforecasts-EpiNow2 model (**?**) made in Germany. A, E: 50% and 90% prediction intervals and observed values for cases and deaths on the natural scale. B, F: Corresponding scores. C, G: Forecasts and observations on the log scale. D, H: Corresponding scores.

6 Human Judgement Forecasting of COVID-19 in the UK

The crowd forecasts submitted to the German and Polish Forecast Hub described in Chapter 4 formed part of an acute COVID-19 response effort and therefore exhibited a few shortcomings that the study presented in this chapter aims to address. In particular, the study in Germany and Poland suffered from a low number of participants, both in terms of the crowd forecast as well as the number of model-based predictions submitted to the Forecast Hub. This makes it difficult to generalise findings. Forecasts were also only evaluated on the natural scale with all the shortcomings discussed in Chapter 5.

This chapter describes a follow-up study conducted in the UK. In order to increase and diversify participation, we organised the study in the form of a public forecasting tournament, the "UK Crowd Forecasting Challenge". This allowed us to analyse whether findings from the initial study would hold in a different setting with a larger pool of participants. Forecasts were analysed both on the natural and on the log scale, providing a more complete picture of the predictive performance of human judgement forecasts.

This chapter also extends the work in Chapter 4 by exploring a novel way to combine human judgement and mathematical modelling as proposed in the original work. Instead of asking forecasters to predict case and death incidences directly, we elicited forecasts of the effective reproduction number R_t which then got mapped to cases and deaths using an epidemiological model. The motivation behind this idea was twofold. On the one hand this might be a possibility to improve forecasts by providing a means to harness the respective strengths of human judgement and mathematical modelling. On the other hand, combining human judgement and mathematical modelling might be a way to make human judgement forecasting more scalable by reducing the cognitive load and the number of forecasts an individual needs to provide.

RESEARCH ARTICLE

REVISED Human judgement forecasting of COVID-19 in the UK [version 2; peer review: 1 approved, 1 approved with reservations]

Nikos I. Bosse[1,2], Sam Abbott[1], Johannes Bracher[3,4], Edwin van Leeuwen[2,5], Anne Cori[6], Sebastian Funk[1,2]

[1]Department of Infectious Disease Epidemiology, London School of Hygiene & Tropical Medicine, London, WC1E 7HT, UK
[2]NIHR Health Protection Research Unit in Modelling & Health Economics, London, UK
[3]Computational Statistics Group, Heidelberg Institute for Theoretical Studies, Heidelberg, Germany
[4]Chair of Statistical Methods and Econometrics, Karlsruhe Institute of Technology, Karlsruhe, Germany
[5]Modelling Economics Unit, UK Health Security Agency, London, UK
[6]MRC Centre for Outbreak Analysis and Modelling, Department of Infectious Disease Epidemiology, School of Public Health, Imperial College London, London, England, UK

v2 **First published:** 19 Sep 2023, **8**:416
https://doi.org/10.12688/wellcomeopenres.19380.1
Latest published: 21 Mar 2024, **8**:416
https://doi.org/10.12688/wellcomeopenres.19380.2

Abstract

Background

In the past, two studies found ensembles of human judgement forecasts of COVID-19 to show predictive performance comparable to ensembles of computational models, at least when predicting case incidences. We present a follow-up to a study conducted in Germany and Poland and investigate a novel joint approach to combine human judgement and epidemiological modelling.

Methods

From May 24th to August 16th 2021, we elicited weekly one to four week ahead forecasts of cases and deaths from COVID-19 in the UK from a crowd of human forecasters. A median ensemble of all forecasts was submitted to the European Forecast Hub. Participants could use two distinct interfaces: in one, forecasters submitted a predictive distribution directly, in the other forecasters instead submitted a forecast of the effective reproduction number *Rt* . This was then used to forecast cases and deaths using simulation methods from the EpiNow2 R package. Forecasts were scored using the weighted interval score on the original forecasts, as well as after

Open Peer Review

Approval Status ? ✓

	1	2
version 2 (revision) 21 Mar 2024		
version 1 19 Sep 2023	? view	✓ view

1. **Daniel J. McDonald**, The University of British Columbia, Vancouver, Canada

2. **Hongru Du**, Johns Hopkins University, Baltimore, USA

Any reports and responses or comments on the article can be found at the end of the article.

applying the natural logarithm to both forecasts and observations.

Results

The ensemble of human forecasters overall performed comparably to the official European Forecast Hub ensemble on both cases and deaths, although results were sensitive to changes in details of the evaluation. *Rt* forecasts performed comparably to direct forecasts on cases, but worse on deaths. Self-identified “experts” tended to be better calibrated than “non-experts” for cases, but not for deaths.

Conclusions

Human judgement forecasts and computational models can produce forecasts of similar quality for infectious disease such as COVID-19. The results of forecast evaluations can change depending on what metrics are chosen and judgement on what does or doesn't constitute a "good" forecast is dependent on the forecast consumer. Combinations of human and computational forecasts hold potential but present real-world challenges that need to be solved.

Keywords
forecasting, human judgement forecasting, COVID-19, UK, United Kingdom, Weighted Interval Score

Corresponding author: Nikos I. Bosse (Nikos.Bosse@lshtm.ac.uk)

Author roles: Bosse NI: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Abbott S**: Conceptualization, Data Curation, Investigation, Methodology, Software, Supervision, Writing – Review & Editing; **Bracher J**: Supervision, Writing – Review & Editing; **van Leeuwen E**: Conceptualization, Supervision, Writing – Review & Editing; **Cori A**: Conceptualization, Supervision, Writing – Review & Editing; **Funk S**: Conceptualization, Funding Acquisition, Project Administration, Supervision, Writing – Review & Editing

Competing interests: NIB is a part-time employee of Metaculus, an online prediction platform.

Grant information: NIB received funding from the National Institute for Health and Care Research (NIHR) Health Protection Research Unit (HPRU) in Modelling and Health Economics (grant code NIHR200908). SA's work was funded by the Wellcome Trust (grant: 210758/Z/18/Z). AC acknowledges funding from the MRC Centre for Global Infectious Disease Analysis (reference MR/ R015600/1) jointly funded by the UK Medical Research Council (MRC) and the UK Foreign, Commonwealth & Development Office (FCDO), under the MRC/FCDO Concordat agreement and is also part of the EDCTP2 programme supported by the European Union; the Academy of Medical Sciences Springboard, funded by the Academy of Medical Sciences, Wellcome Trust, the Department for Business, Energy and Industrial Strategy, the British Heart Foundation, and Diabetes UK (reference SBF005\1044); and the National Institute for Health and Care Research (NIHR) Health Protection Research Unit in Modelling and Health Economics, a partnership between the UK Health Security Agency, Imperial College London and LSHTM (grant code NIHR200908). EvL acknowledges funding by the National Institute for Health and Care Research (NIHR) Health Protection Research Unit (HPRU) in Modelling and Health Economics (grant number NIHR200908) and the European Union's Horizon 2020 research and innovation programme - project EpiPose (101003688). The work of JB was supported by the Helmholtz Information and Data Science Project SIMCARD as well as Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 512483310. SF's work was supported by the Wellcome Trust (grant: 210758/Z/18/Z) and the HPRU (grant code NIHR200908). The views expressed are those of the authors and not necessarily those of the UK Department of Health and Social Care (DHSC), NIHR, or UKHSA.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Bosse NI, Abbott S, Bracher J *et al.* **Human judgement forecasting of COVID-19 in the UK [version 2; peer review: 1 approved, 1 approved with reservations]** Wellcome Open Research 2024, **8**:416
https://doi.org/10.12688/wellcomeopenres.19380.2

First published: 19 Sep 2023, **8**:416 https://doi.org/10.12688/wellcomeopenres.19380.1

REVISED **Amendments from Version 1**

We added a more detailed contextualisation of the study period (May to September 2021) and an explanation of the various factors that contributed to the pattern of observed cases and deaths from COVID-19 in the UK at that time. We also added a new Figure to illustrate the study period. We included details on the study authors who made forecasts as participants. We clarified parts of the discussion related to the evolution of the case fatality ratio (CFR) over the study period and provided references. We also clarified that our human forecasts were included in the overall Hub ensemble against which they are compared, likely leading us to underestimate the differences between the two. We added suggestions for further research, for example on priming effects from defaults shown in the user interface or on the effect that the availability of additional qualitative data might have on forecast accuracy.

Any further responses from the reviewers can be found at the end of the article

Introduction

Infectious disease modelling and forecasting has attracted wide-spread attention during the COVID-19 pandemic and helped inform decision making in public health organisations and governments[1,2]. Most forecasts used to inform decision making were based on computational models of COVID-19, but some authors also explored human judgement forecasting as an alternative or in combination[3–6].

Past research found that in the context of infectious disease forecasting, human judgement forecasts could achieve predictive performance broadly comparable to forecasts generated based on mathematical modelling, in particular when forecasting incident cases, rather than lagged indicators indicators like deaths. Farrow *et al.*[7] found that an aggregate of human predictions outperformed computational models when predicting the 2014/15 and 2015/16 flu season in the US. However, a comparable approach performed worse than computational models at predicting the 2014/15 outbreak of chikungunya in the Americas. Bosse *et al.*[3] found an ensemble of human forecasters to outperform an ensemble of computational models when predicting cases of COVID-19 in Germany and Poland, but performing worse when predicting incident deaths. Similarly, McAndrew *et al.*[5] reported an ensemble of human forecasters to perform comparably to an ensemble of computational models when predicting incident COVID-19 cases, and worse when predicting incident deaths. Farrow *et al.*[7] and in particular Bosse *et al.*[3] struggled to recruit many participants (numbers of active forecasters ranged from 22 to 61 in McAndrew *et al.*[5], 7 to 24 in Farrow *et al.*[7], and 4 to 10 in Bosse *et al.*[3]). It is important to note that in previous studies (and also this one) human forecasters were free to use any resources, including computational models, in the process of creating a forecast, making it difficult to completely separate human judgement and computational modelling.

In some situations, human judgement forecasting may have advantages relative to computational models. Human judgment may be particularly useful to provide timely forecasts in situations where data is sparse and many parameters are hard to quantify. Humans are also generally able to answer a broad set of question (such as for example the likelihood that a given actor will take some specified action) and can take factors into account that are hard to encode in a computational model. On the other hand, human judgement forecasting is difficult to scale due to the time and effort required, and humans may be at a disadvantage at tasks that strongly benefit from the ability to perform complex computations. Also, the use of human judgement forecasts by decision makers may be complicated by the lack of clarity of the basis on which they were made.

Methods that aim to combine human judgement and mathematical modelling are therefore appealing, though we note that presenting this as a binary choice is misleading. Most computational models in use in epidemiology have at least some element of human judgement supporting their structure or usage. Also, human forecasters often make use of approaches such as calculating a base rate of incidences, or extrapolating current trends, which are in reality equivalent to simple models. One explicit method to combine separate human judgement and computational model forecasts with the goal of improving predictive performance is an ensemble. This has been shown to improve performance across model types[5]. Farrow *et al.*[7], Bosse *et al.*[3], Swallow *et al.*[8] and others suggested additional possibilities in the context of infectious diseases that may also help reduce the amount of human effort required. One approach is to use human forecasts, for example of relevant disease parameters, as an input to computational modelling. Another approach is to use mathematical modelling in explicit combination with human judgement, for example by giving experts the option to make post-hoc adjustments to model outputs. Bosse *et al.*[3] proposed asking human forecasters to forecast the effective reproduction number R_t (the average number of people an infected person would infect in turn) based on modelled estimates and to then use this forecast in a mathematical simulation model in order to obtain forecasts for observed case and death numbers.

This paper represents a follow-up study to Bosse *et al.*[3] in the United Kingdom with one- to four-week ahead forecasts made over the course of thirteen weeks between May 24 and August 16, 2021. The study period is after the second wave of COVID-19 in the UK (which peaked in January 2021) and falls into a time when restrictions in the UK were gradually lifted as part of the roadmap out of lockdown (with final restrictions lifted on July 19, 2021). Forecasts were elicited from experts and laypeople as part of a public forecasting tournament, the "UK Crowd Forecasting Challenge", using a web application. All forecasts were submitted to the European COVID-19 Forecast Hub, one of several Forecast Hubs that have been systematically collating forecasts of different COVID-19 forecast targets in the US[1], Germany and Poland[9,10], and Europe[11]. This study aims to investigate whether the original findings in Bosse *et al.*[3] with respect to forecaster performance replicate in a different country, in a different

time period, and with an increased number of participants. In addition, it explores the approach proposed in Bosse *et al.*[3] to ask participants for a forecast of the estimated effective reproduction number R_t which is then translated into a forecast of cases and deaths using a simulation model. We describe this approach as human in the loop computational modelling and consider it a formalisation of often practiced manual intervention in computational forecasts.

Methods

Interaction with the European Forecast Hub

The European COVID-19 Forecast Hub[11] was launched in March 2021 in order to elicit weekly predictions for various COVID-19 related forecast targets from different research groups. The forecasts evaluated in this study were submitted every Monday before 11.59pm GMT between May 24 2021 and August 16 2021. Forecasts were made for incident weekly reported numbers of cases of and deaths from COVID-19 on a national level for various European countries over a one to four week forecast horizon. While forecasts were submitted on Mondays, weeks were defined as epidemiological weeks, ending on a Saturday, and starting on Sunday. Forecast horizons were therefore in fact 5, 12, 19 and 26 days. Submissions to the European Forecast Hub followed a quantile-based format with 23 quantiles of each output measure at levels 0.01, 0.025, 0.05, 0.10, 0.15,. . . , 0.95, 0.975, 0.99. Every week, forecasts submitted to the hub were automatically checked for conformity with the required format and all eligible forecasts combined into different ensembles. Until the 12th of July 2021 the default Hub ensemble ("EuroCOVIDhub-ensemble") shown on all official Forecast Hub visualisations (https://covid19forecasthub.eu/) was a mean ensemble (i.e., the α-quantile of the ensemble is given by the mean of all submitted α-quantiles). From the 29th of July onwards, the default Forecast Hub ensemble became a median ensemble. The median number of models included in the Forecast Hub ensemble for the UK during the study period was 9 for cases and 10 for deaths (see Figure SI.1 in the SI).

Ground-truth data on daily reported test positive cases and deaths linked to COVID-19 were provided by the European Forecast Hub and sourced from the Johns Hopkins University (JHU). Data were subject to reporting artifacts and revisions. All data points were marked as anomalous retrospectively by the European Forecast Hub if in subsequent updates data was changed by more than 5 percent. In August 2022 JHU switched the data source for their UK death numbers from "deaths within 28 days of a positive COVID test" to "Deaths with COVID-19 on the death certificate" and revised all their past data to guarantee consistency. The 2021 UK ground truth death data as it was made available through the European Forecast Hub in 2021 is therefore substantially different and on average lower than the data available as of early 2023. Data revisions are displayed in Figure SI.2 in the Supplementary Information[12]. All results presented here were derived based on the original data available in 2021, which were available through the European COVID-19 Forecast Hub GitHub repository (https://github.com/covid19-forecast-hub-europe/covid19-forecast-hub-europe).

Human judgement forecasts

Forecasts of incident cases and deaths linked to COVID-19 in the UK were elicited from individual participants every week through a web application (https://cmmid-lshtm.shinyapps.io/crowd-forecast/) described in 3. The application is based on `R`[13] `shiny`[14] and is available as an `R` package called `crowdforecastr`[15]. When signing up, participants could self-identify as "experts" if they worked in infectious disease modelling or had professional experience in any related field.

The web application offered participants two different ways of making a forecast, called 'direct' (or 'classical') and 'R_t forecast'. To make a 'direct' forecast (as described in more detail in 3), participants selected a predictive distribution (by default a log-normal distribution) and adjusted the median and width of the distribution to change the central estimate and uncertainty at each forecast horizon.

Just as in the previous study, the default forecast shown was a repetition of the last known observation with constant uncertainty around it. The shown distribution was the exponential of a normal distribution with mean log(last value) and uncertainty set to the standard deviation of the last four changes in weekly log observed forecasts (i.e., as σ(log(value4) − log(value3), log(value3) − log(value2), . . .)). In addition to information about past observations, participants could see various metrics and data such as the test positivity rate and vaccination rate sourced from Our World in Data[16]. Figure SI.3 in the Supplementary Information[12] shows a screenshot of the forecast interface for direct forecasts.

In addition to the 'direct' forecasts, we implemented a second forecasting method ('R_t forecasts'), where we asked participants to make a forecast of the effective reproduction number R_t. This forecast was made based on a baseline estimate produced by the `EpiNow2`[17] `R`[13] package effective reproduction number model which we also used in 3 as a standalone computational model. The estimate produced by `EpiNow2` was shown as the default forecast and could be adjusted by the user. The resulting R_t forecast was then translated into a forecast of cases using the simulation model from the `EpiNow2 R` package, which implements a renewal equation based[18] generative process for latent infections. We chose a Gaussian Process prior with mean 0 for the first differences of the effective reproduction number in time, implying that in the absence of informative data the reproduction number would remain constant on average, with uncertainty increasing with the temporal distance to informative data points. Latent infections were convolved with delay distributions representing the incubation period and reporting delay, and assumed to follow a negative binomial observation model with a day of the week effect to produce an estimate of reported cases. This approach has been widely used for short-term forecasting[3,11] and used to produce reproduction number estimates[19–21]. Further details are given in the Supplementary Information[12].

To obtain forecasts for deaths, we similarly fit a model that convolved observed and predicted reported cases as implied by the R_t forecast over a delay distributions[20,21] and scaled

them by a fixed ratio to model the time between a case report and a reported death and the case fatality ratio using the `EpiNow2 R` package[17]. Further details are given in the Supplementary Information[12].

As R_t-estimates up to at least two weeks prior to the forecast data were uncertain due to their dependence on partially complete observations of underlying infections given the delays from infection to report, we also asked participants to submit an estimate of R_t for the two weeks prior to the current forecast date. Participants were therefore asked to estimate/predict six R_t values, four of them beyond the forecast horizon. In order to obtain sample trajectories needed as input for the simulation model, we drew 1000 samples from the six provided distributions. These samples were ordered and corresponding samples treated as one sample trajectory. Samples for daily values were obtained by linearly interpolating between weekly samples.

Upon pressing a button, participants could see a preview of the evolution of cases implied by their current R_t forecast. However, due to lack of development time, participants could not preview the death forecast implied by their current input for R_t nor could they influence the estimated case fatality ratio or delay between reported cases and reported deaths. Figure SI.4 in the Supplementary Information[12] shows a screenshot of the forecast interface for R_t forecasts.

Every week, we submitted an ensemble of individual forecasts to the European Forecast Hub. In contrast to the ensemble of human forecasts described in Bosse *et al.*[3], we used the quantile-wise median, rather than the quantile-wise mean to combine predictions, drawing upon insights gained from the COVID-19 Forecast Hubs[22]. We submitted three different ensembles to the Hub: The first one, "epiforecasts-EpiExpert_direct" (here called "direct crowd forecast" or "crowd-direct") was a quantile-wise median ensemble of all the direct forecasts. "epiforecasts-EpiExpert_Rt" (here called "R_t forecast" or "crowd-rt") was a median ensemble of all forecasts made through the R_t interface. "epiforecasts-EpiExpert" (here called "combined crowd ensemble" or "crowd-ensemble") was a median ensemble of all forecasts together. A participant could enter the combined crowd ensemble twice if they had submitted both a direct and an R_t forecast. Before creating the ensemble, we deleted forecasts that were clearly the result of a user or software error (such as forecasts that were zero everywhere). Our combined crowd ensemble, "epiforecasts-EpiExpert", but not the other two, entered the official European COVID-19 Forecast Hub ensemble ("EuroCOVIDhub-ensemble").

The UK Crowd Forecasting Challenge

To boost participation compared to our last crowd forecasting study in Germany and Poland[7] which struggled in this regard, we announced an official tournament, the "UK Crowd Forecasting Challenge". Participants were asked to submit weekly predictions for reported cases and deaths linked to COVID-19 in the United Kingdom one to four weeks into the future. Everyone who had submitted a forecast for targets in the UK during the tournament period from the 24th of May 2021 to the 16th of August 2021 was deemed a participant and eligible for a prize. The first prize was 100 GBP, second prize 50 GBP and third prize 25 GBP. Participant performance was determined using the mean weighted interval score (WIS) on the log scale (see details in the next Section), averaged across forecast dates, horizons and forecast targets. For the tournament ranking, participants who did not submit a forecast in a given week were assigned the median score of all other participants who submitted a forecast that week. The UK crowd forecasting challenge was announced over Twitter and our networks. In addition, we created a project website, https://crowdforecastr.org, made weekly posts on Twitter and sent participants who had registered on the online application weekly emails with a reminder and a summary of their past performance. A public leaderboard was available on our website https://epiforecasts.io. Participants could choose to make a direct forecast as well as an R_t forecast and were counted as two separate forecasters and eligible for prizes twice. Weekly forecasts had to be submitted between Sunday 12pm and Monday 8pm UK time.

Analysis

We scored forecasts using the weighted interval score[23]. For (1-α)·100% prediction interval, the interval score is computed as

$$IS_\alpha(F,y) = (u-l) + \frac{2}{\alpha}\cdot(l-y)\cdot 1(y \le l) + \frac{2}{\alpha}\cdot(y-u)\cdot 1(y \ge u),$$

where 1() is the indicator function, y is the true value, and l and u are the $\frac{\alpha}{2}$ and $1-\frac{\alpha}{2}$ quantiles of the predictive distribution F, i.e., the lower and upper bound of a single prediction interval. For a set of K prediction intervals and the median m, the score is computed as a weighted sum,

$$WIS = \frac{1}{K+0.5}\cdot\left(w_0\cdot|y-m| + \sum_{k=1}^{K} w_k \cdot IS_\alpha(F,y)\right),$$

where w_k is a weight for every interval. Usually, $w_k = \frac{\alpha_k}{2}$ and $w_0 = 0.5$.

The WIS is a strictly proper scoring rule yielding non-negative values, with smaller values implying better performance. A forecaster, in expectation, optimises their score by providing a predictive distribution F that is equal to the data-generating distribution G, and is therefore incentivised to report their true belief. The WIS can be understood as an approximation of the continuous ranked probability score (CRPS, Gneiting *et al.*[24]) for forecasts in a quantile-based format. The CRPS, in turn, represents a generalisation of the absolute error to predictive distributions. The WIS can be decomposed into three separate penalty components (corresponding to the three terms in the definition of the interval score): forecast dispersion (i.e., uncertainty of forecasts), overprediction and underprediction.

Bosse *et al.*[25] recently suggested to transform forecasts and observations using the natural logarithm prior to applying the WIS to better reflect the exponential nature of the underlying

disease process. We, therefore, also compute WIS values after transforming all forecasts and observations using the function $f : x \to \log(x + 1)$. In the following, we refer to WIS scores obtained without a transformation as "scores on the natural scale", and WIS values obtained after log-transforming forecasts and observations as "scores on the log scale". To make scores easier to interpret, we report relative WIS scores, where the average score for a given model was divided by the average score for the European Forecast Hub ensemble ("EuroCOVIDhub-ensemble"). In addition, we computed ranks based on WIS values.

In order to measure probabilistic calibration[24], we used the empirical coverage of all central 50% and 90% prediction intervals. Empirical coverage refers to the percentage of observations falling inside any given central prediction interval (e.g., the cumulative percentage of observed values that fall inside all central 50% prediction intervals).

If not otherwise stated, we present results for two-week-ahead forecasts, following the practice adopted by the COVID-19 Forecast Hubs, which found predictive performance to be poor and unreliable beyond this horizon[1,9,11]. We analysed all forecasts stratified by forecast target (cases or deaths), forecast horizon, and forecast approach. We compared the performance of the direct vs. R_t forecasting approach using instances where we had both a direct forecasts and an R_t forecast from the same person.

For self-reported "experts" and "non-experts", a simple comparison of scores would be confounded by individual differences in participation and the timing of individual forecasts. We therefore compared the performance of self-reported "experts" vs. "non-experts" by creating and evaluating two modified median ensembles, one including only "experts" and the other only "non-experts".

Forecasts were evaluated using the `scoringutils`[26] package in `R`. All code and data used for this analysis, including individual-level forecasting data is available at https://github.com/epiforecasts/uk-crowd-forecasting-challenge. All code used to submit the forecasts to the European Forecast Hub is available at https://github.com/epiforecasts/europe-covid-forecast.

Ethics statement
This study has been approved by the London School of Hygiene & Tropical Medicine Research Ethics Committee (reference number 22290). Consent from participants was obtained in written form.

Results
Observed values
The study period (forecasts were made between May 24 and August 16, 2021, for targets between May 29 and September 11, 2021) was characterised by an increase in the number of cases and deaths in the United Kingdom. Reported cases in particular rose rapidly compared to pre-study levels, with a peak on July 17, 2021, followed by a trough and another subsequent increase in numbers. Death numbers remained almost constant in the first four weeks of the study period, followed by a steady increase until the end of the study period in September 2021. This increase in the case and death numbers coincides with the rise of the Delta variant in the UK at the beginning of May[27,28] as well as the European Football Championship[29]. Reported cases were likely influenced by an increased uptake of the NHS COVID-19 app in spring and summer 2021[30]. An overview of the reported case and death numbers is shown in Figure 1.

Crowd forecast participation
A total number of 90 participants submitted forecasts (more precisely, forecasts were submitted from 90 different accounts, some of them anonymous). Out of 90 participants, 21 self-identified as "experts", i.e., stated they had professional experience in infectious disease modelling or a related field.

The median number of unique participants in any given week was 17, the minimum was 6 and the maximum was 51. This was higher than the number of participants in 3 (which had a median number of 6, a minimum of 2, and a maximum 10). With respect to the number of submissions from an individual participant, we observed similar patterns as 3: An individual forecaster participated on average in 2.6 weeks out of 13. The median number of submissions from a single individual was one, meaning that similar to 3 most forecasters dropped out after their first submission. Only five participants submitted a forecast in ten or more weeks and only two submitted a forecast in all thirteen weeks, one of whom is an author on this study (S. Abbott). Three other authors participated in the study (S. Funk, N. Bosse, and E. van Leuwwen). A total of 535 forecasts were submitted by human forecasters, 118 (22%) of these were submitted by authors of this study. The number of direct forecasts (median: 13 for cases and 12 for deaths) was higher than the number of R_t forecasts (median: 6 for both cases and deaths) in all weeks (see Figure 2A). The median number of "non-experts" (11 for cases, 10 for deaths) was higher than the median number of "experts" (8 for cases and deaths) (see Figure 2B).

Case forecasts
At the beginning of the study period, human forecasters as well as the Forecast Hub ensemble, consistently underpredicted case numbers (see Figure 5A). All forecasting approaches overshot the peak in case numbers on July 17, 2021, overpredicting case numbers severely in the three weeks after, followed again by a small tendency to underpredict when case numbers rose once more in the 4th week after the peak.

All forecasting approaches exhibited underdispersion when predicting cases, meaning that forecasts on average were too narrow and not uncertain enough. Empirical coverage for case forecasts was below nominal coverage for all forecasting approaches for forecasts more than one week into the future (see Figure 3E,F). For 50% prediction intervals, empirical coverage was worst for the direct crowd forecasts (0.31), best for the R_t forecasts (0.46) and in between for the Hub ensemble and the crowd ensemble (both 0.38, see Table 1). For 90%

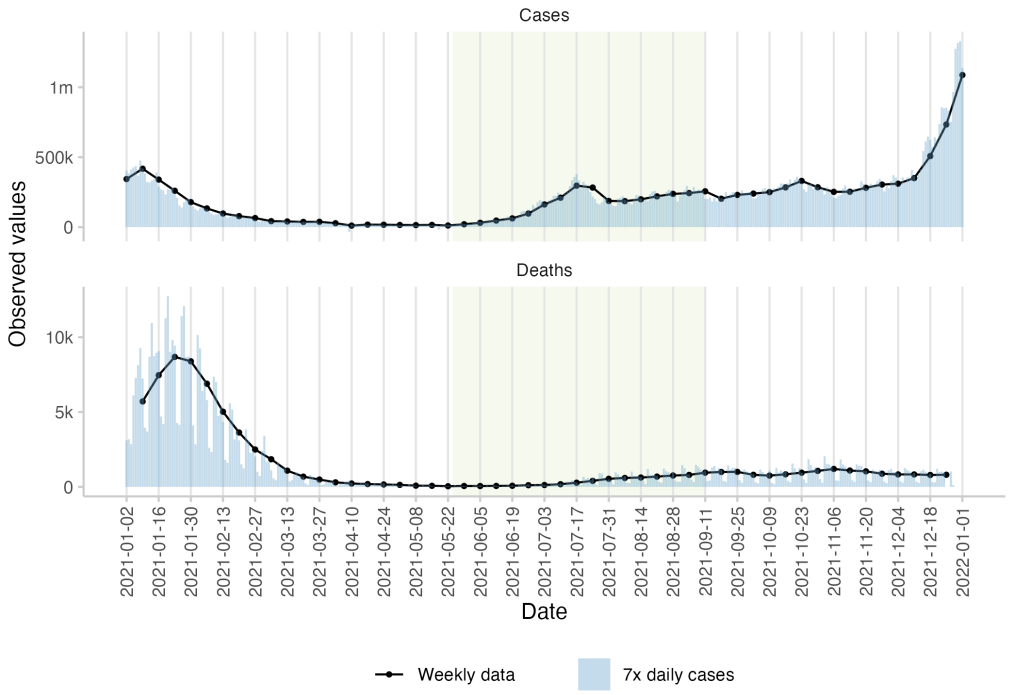

Figure 1. Observed cases and deaths of COVID-19 in the UK. Observed daily (bars) and weekly (black lines and points) numbers of cases and deaths as available through the European Forecast Hub when the study concluded in 2021. The green rectangle marks the study period from May 24 until September 11, 2021. Daily numbers were multiplied by seven in order to appear on the same scale as weekly numbers.

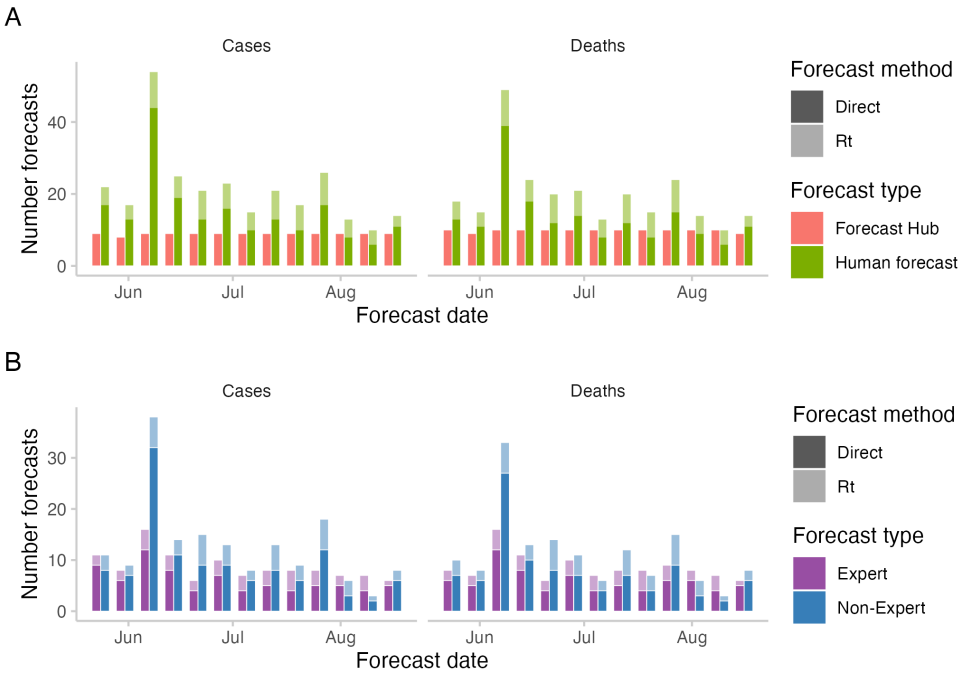

Figure 2. Number of forecasts across the study period. A: number of forecasts included in the Hub ensemble and the combined crowd ensemble. **B**: number of forecasts by "experts" and "non-experts". Expert status was determined based on the participant's answer to the question whether they "worked in infectious disease modelling or had professional experience in any related field".

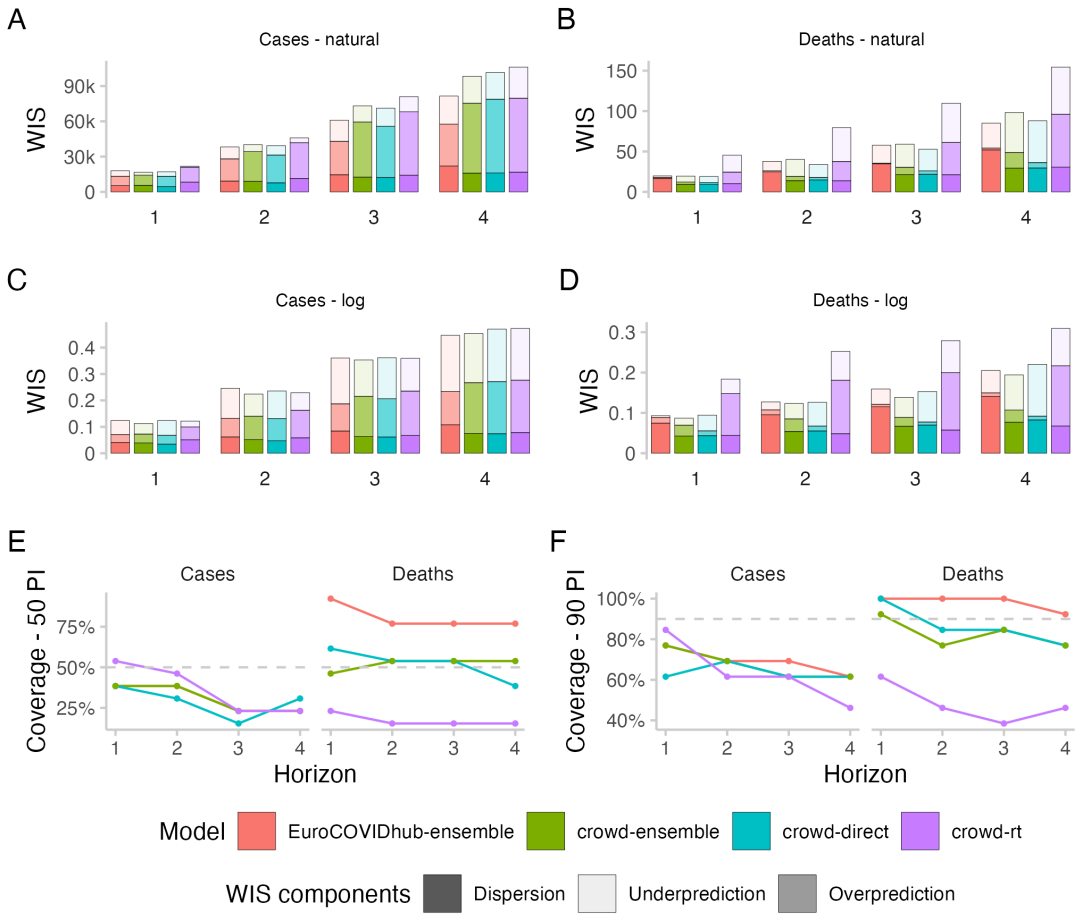

Figure 3. Predictive performance across forecast horizons. A–**D**: WIS stratified by forecast horizon for cases and deaths on the natural and log scale. **E**, **F**: Empirical coverage of the 50% and 90% prediction intervals stratified by forecast horizon and target type. Grey dashed lines denote the nominal coverage that a model should ideally achieve.

Table 1. Performance for two-week-ahead forecasts. Values have been cut to three significant digits and rounded.

Model	Target	WIS - natural			WIS - log scale			Coverage 50%	Coverage 90%
		abs.	rel.	sd	abs.	rel.	sd		
EuroCOVIDhub-ensemble	Cases	38.2k	1	55.6k	0.25	1	0.22	0.38	0.69
crowd-ensemble	Cases	40.1k	1.05	69.4k	0.22	0.91	0.25	0.38	0.69
crowd-direct	Cases	39.3k	1.03	67k	0.23	0.96	0.27	0.31	0.69
crowd-rt	Cases	45.9k	1.2	74.7k	0.23	0.93	0.24	0.46	0.62
EuroCOVIDhub-ensemble	Deaths	37.9	1	26.9	0.13	1	0.04	0.77	1
crowd-ensemble	Deaths	40.2	1.06	41.5	0.12	0.97	0.07	0.54	0.77
crowd-direct	Deaths	33.9	0.89	30.6	0.13	0.99	0.08	0.54	0.85
crowd-rt	Deaths	79.5	2.1	72.7	0.25	1.98	0.13	0.15	0.46

prediction intervals, coverage was worst for the R_t forecasts (0.62) and slightly better for the other approaches (all 0.69). Coverage for all forecasts deteriorated further with increasing forecast horizon (see Figure 3E,F).

In terms of WIS on the log scale, all human forecasting approaches outperformed the Forecast Hub ensemble for two week ahead forecasts of cases (see Figure 3). WIS values relative to the Hub ensemble (=1) were 0.91 for the combined crowd ensemble, 0.96 for the direct crowd forecasts and 0.93 for the R_t forecasts (see Table 1). In contrast, in terms of WIS on the natural scale, the Hub ensemble outperformed all human forecasting approaches. Relative WIS values on the natural scale for two week ahead forecasts were 1.05 for the combined crowd ensemble, 1.03 for the direct crowd forecasts and 1.2 for the R_t forecasts. The discrepancy between performance on the log and natural scale can be attributed to case forecasts from the Hub ensemble tending to be lower than forecasts from human judgement approaches (see Figure 4). On the natural scale, this resulted in smaller overprediction penalties, putting it ahead of human forecasts (see Figure 3A,C). On the log scale, however, it led to large penalties for underprediction.

Performance of the Hub ensemble relative to the human forecasting approaches improved with increasing forecast horizon (see Figure 3). For a four-week-ahead forecast horizon, the Hub ensemble outperformed all other approaches both on the log scale (rel. WIS values the human forecasts of 1.02, 1.05, 1.06) and on the natural scale (rel. WIS values of 1.21, 1.25, 1.3) (compare Table SI.1 in the Supplementary Information[12]).

In terms of relative model ranks for two week ahead forecasts, the Hub ensemble and the R_t forecast showed a higher variance than the combined crowd ensemble and the direct forecasts (See Figure 5), despite forecasts being about the same or more dispersed (see Figure 3). Both the Hub ensemble and the R_t forecast were more often in first place than other approaches (4 times each, both on the log and on the natural scale). However, they were also most often in the last place (Hub ensemble: 6 on the log scale and 5 on the natural scale, R_t : 5 on the log scale and 6 on the natural scale). The direct forecasts placed relatively equally in places 1-4. The crowd ensemble never placed fourth, but also had the lowest number of first places (2, both on the log and the natural scale). Aggregated model ranks only changed marginally when switching between the log and the natural scale (see Figure 5).

When comparing WIS values on the log scale with those on the natural scale, scores were more equally distributed across the study period on the log scale and more weight was given to forecasts in June and July which underpredicted the extent to which case number would rise (see Figure 4). On the natural scale, the WIS as a measure of the absolute distance between forecast and observation increased or decreased with the magnitude of the forecast target[23,25]. Average scores were therefore dominated by performance around the peak when cases were highest, in particular by forecasts made on the 19th of July for the 31st of July (see Figure 4). For all forecasting approaches, overprediction was the largest contributor to overall scores (see Figure 3A). On the log scale, underprediction played a larger role (see Figure 3C). Switching between scores on the log and on the natural scale had the strongest effect on the R_t forecasts, which had a relative WIS value of 0.96 on the log scale and 1.2 on the natural scale. The R_t forecasts tended to be higher than both the direct forecasts and the Forecast Hub ensemble, especially around the peak, leading to high scores on the natural scale, but not on the log scale.

Death forecasts

In the first part of the study period, most forecasting approaches (albeit not the direct crowd forecasts), showed a tendency to overpredict the increase in death numbers (see Figure 5B). All forecasting approaches started to underpredict death numbers four weeks after the peak in case numbers on July 17, 2021, expecting a consequent drop in deaths that did not occur.

All forecasting approaches except the R_t forecasts showed higher empirical coverage for deaths than for cases (see Figure 3). Forecasts from the Hub ensemble generally tended to be wider than the human forecasts (see Figure 4 and Figure 3B,D). For 50% prediction intervals, the Hub ensemble exceeded the nominal coverage noticeably (0.77) (see Table 1). R_t forecasts failed to get close to nominal coverage (0.15), while the combined crowd ensemble and the direct forecasts had empirical coverage close to nominal coverage (both 0.54). For 90% prediction intervals, the Hub ensemble again exceeded nominal coverage and covered all observations (1) while the R_t forecasts again failed to get close to nominal coverage (0.46). The crowd ensemble exhibited some underdispersion (0.77) while the direct forcecasts almost reached nominal coverage for two week ahead forecasts of deaths (0.85).

In terms of WIS on the log scale for two week ahead predictions of deaths, the combined crowd ensemble (0.97) and the direct crowd forecasts (0.99) were marginally ahead of the Hub ensemble, while the R_t forecasts performed noticeably worse (1.98) (see Figure 3D and Table 1). For the Hub ensemble, the dispersion component played by far the largest role, while this was less the case for the human forecasts, which got higher penalties from both over- and underprediction. Combining the R_t forecasts and the direct forecasts led to an ensemble that performed better than either of them alone on the log scale despite the poor overall performance of the R_t forecasts. In terms of WIS on the natural scale, only the direct forecasts (0.89) performed better for two week ahead death predictions than the Hub ensemble, while the combined crowd ensemble performed slightly worse (1.06) and the R_t forecasts again noticeably worse (2.1).

In terms of relative model ranks for two week ahead death forecasts, the R_t forecasts took the fourth place most often (9

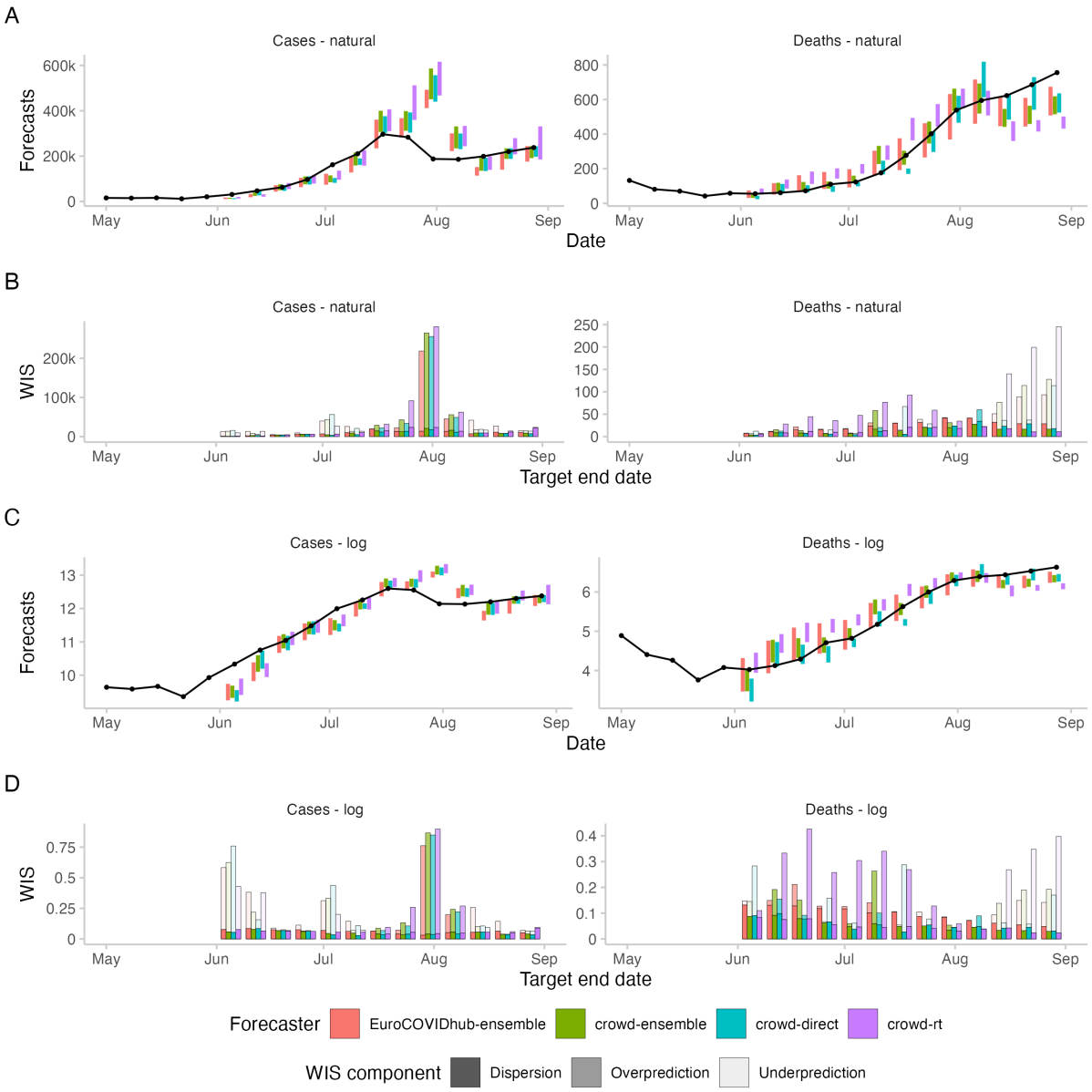

Figure 4. Forecasts and corresponding WIS for 2-week ahead forecasts of cases and deaths from COVID-19 in the UK. A: 50% prediction intervals (coloured bars) and observed values (black line and points) for cases and deaths on the natural scale. **B**: Corresponding WIS values, decomposed into dispersion, overprediction and underprediction. **C**: 50% prediction intervals on the log scale, i.e., after applying the natural logarithm to all forecasts and observations. **D**: Corresponding WIS on the log scale, i.e., the WIS applied to the log-transformed forecasts and observations.

on the log scale and 10 on the natural scale), while the direct forecasts placed first most often (5 on the log scale and 6 on the natural scale, see Figure 5). Again, the crowd ensemble never placed fourth.

When comparing scores on the log and on the natural scale, scores on the log scale were again more evenly distributed across the study period. On the natural scale, high scores were concentrated around the end of the study period, when death incidences were highest (see Figure 4).

R_t forecasts

For cases, where participants could observe the case forecast implied by their R_t forecast, predictive performance was similar between corresponding direct and R_t forecasts for most forecasters who had submitted both (see Figure 6). For

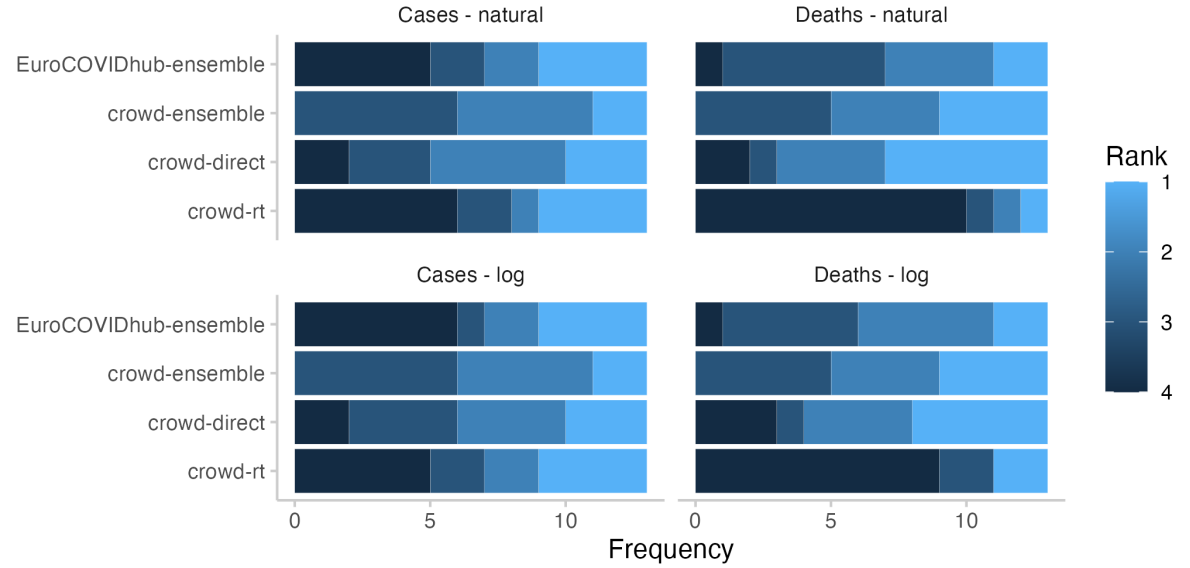

Figure 5. Ranks for all forecasting approaches for two week ahead forecasts. Colours indicate how often (out of 13 forecasts) a given approach got 1st, 2nd, 3rd, or 4th rank.

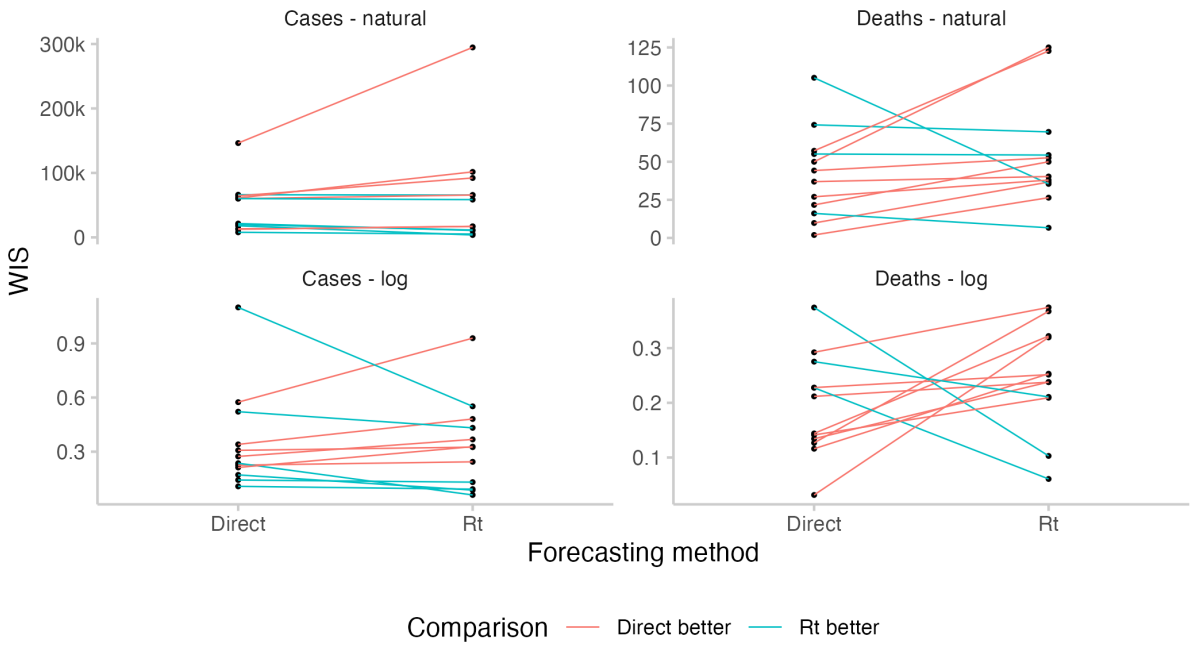

Figure 6. Comparison of predictive performance of individual forecasters using either the direct forecasting or R_t interface. Comparisons are based only on those instances where forecasters have submitted a prediction using both interfaces. The absolute level for a given forecaster relative to others is not meaningful as forecasters differ in the amounts of forecasts they have submitted and when.

deaths, where forecasters could not see the incidence forecast implied by their R_t forecast or manually adjust the case fatality rate, performance of the R_t forecasts was significantly worse. From June to the end of July, R_t forecasts overpredicted deaths and were noticeable higher than other forecasts, whereas in August, R_t forecasts underpredicted deaths and were substantially lower than other forecasts (see Figure 4). In particular, R_t forecasts for deaths were worse than the corresponding

direct death forecasts for most forecasters (see Figure 6). Changing from the direct forecasting method to R_t forecasting for cases tended to improve scores for better forecasters and decrease scores for worse forecasters, although sample sizes and the size of the observed effect are both small.

Combining direct crowd forecasts and R_t forecasts improved performance on the log scale compared to both direct and R_t forecasts alone across all horizons and target types. This was not the case on the natural scale, where direct forecasts performed better than the R_t and the direct forecasts for both cases and deaths across most horizons. Only for case forecasts four weeks ahead on the natural scale was the combined ensemble better than the direct forecasts. However, even on the natural scale, performance of the combined ensemble was better than the average of the WIS of direct and R_t forecasts.

Experts and non-experts

A median ensemble of two week ahead forecasts restricted to only those made by either "experts" or "non-experts" (determined based on self-reported experience in infectious disease modelling or a related field) performed worse than the combined crowd example, both for cases and deaths and both on the log scale and on the natural scale (see Figure 7 and Table 2 and Figure 2B for a visualisation of participation). The median number of "non-experts" was 11 for cases and 10 for deaths, which was higher than the median number of "experts", which was 8 for cases and deaths.

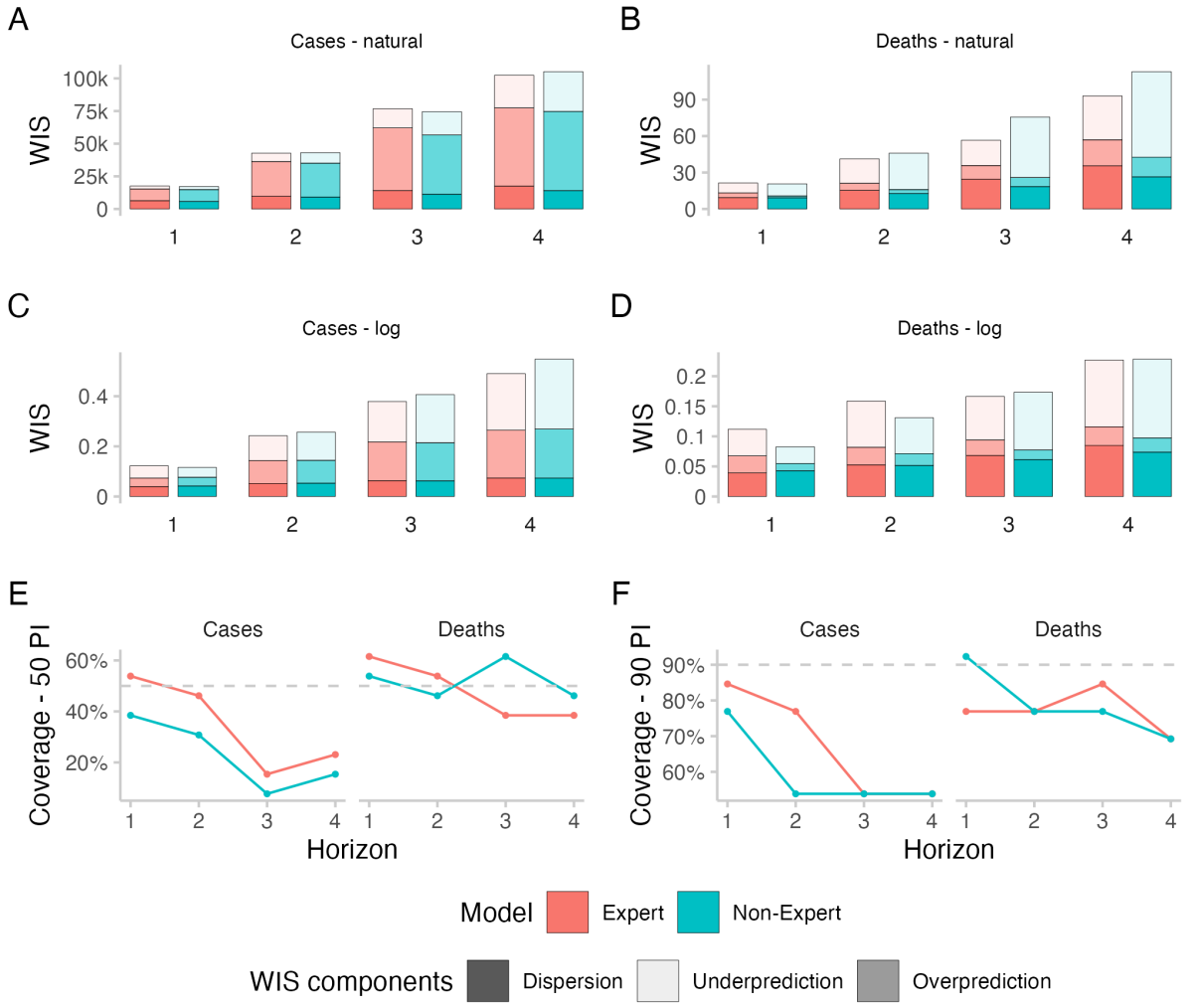

Figure 7. Predictive performance of self-reported "experts" and "non-experts" across forecast horizons. Forecasts from "experts" and "non-experts" were combined to two separate median ensembles, including both direct and R_t forecasts. **A**–**D**: WIS stratified by forecast horizon for cases and deaths on the natural and log scale. **E**, **F**: Empirical coverage of the 50% and 90% prediction intervals stratified by forecast horizon and target type. Grey dashed lines denote the nominal coverage that a model should ideally achieve.

Table 2. Performance for two-week-ahead forecasts of experts and non-experts. Values have been cut to three significant digits and rounded.

Model	Target	WIS - natural			WIS - log scale			Coverage 50%	Coverage 90%
		abs.	rel.	sd	abs.	rel.	sd		
crowd-ensemble	Cases	40.1k	1	69.4k	0.22	1	0.25	0.38	0.69
Expert	Cases	42.7k	1.06	74.9k	0.24	1.08	0.28	0.46	0.77
Non-Expert	Cases	43.1k	1.07	67k	0.26	1.14	0.25	0.31	0.54
crowd-ensemble	Deaths	40.2	1	41.5	0.12	1	0.07	0.54	0.77
Expert	Deaths	41.2	1.03	41.8	0.16	1.29	0.15	0.54	0.77
Non-Expert	Deaths	45.9	1.14	56.8	0.13	1.06	0.08	0.46	0.77

When comparing two week ahead forecasts from "experts" and "non-experts", the ensemble of "experts" was better calibrated (see Figure 7). For cases, "experts" achieved better scores than "non-experts" both on the log and on the natural scale. WIS values *relative to the combined crowd ensemble* were 1.08 for "experts" and 1.14 for "non-experts" on the log scale and 1.06 for "experts" and 1.07 for "non-experts" on the natural scale (see Table 2). For deaths, "experts" performed worse than "non-experts" in terms of WIS on the log scale (WIS relative to the combined crowd ensemble: 1.29 vs. 1.06), but better on the natural scale (1.03 vs. 1.14). Both the "expert"- and the "non-expert"-ensemble had similar proportions of R_t forecasts (mean of 32% for "experts" and 32.2% for "non-experts" across cases and deaths together).

For four weeks ahead forecasts of cases, the combined ensemble outperformed both "experts" and "non-experts" on the log scale as well as on the natural scale. "Experts" performed better than "non-experts" both on the log scale (WIS value relative to the combined crowd ensemble of 1.08 for "experts" vs. 1.21 for "non-experts") and on the natural scale (1.04 vs. 1.07). For four week ahead forecasts of deaths, "Experts" performed better than "Non-experts" on the log scale (1.17 vs. 1.18) as well as on the natural scale (0.95 vs. 1.15).

Discussion

In this paper, we presented a follow-up study to Bosse *et al.*[3], analysing human judgement forecasts of cases of and deaths from COVID-19 in the United Kingdom submitted to the European COVID-19 Forecast Hub between the 24th of May and the 16th of August 2021. Human judgement forecasts were generated using two different forecasting approaches, a) direct forecasts of cases and deaths and b) forecasts of the effective reproduction number R_t, which were based on estimates from an open source effective reproduction number estimation model and also relied on this model, along with a second model relating cases and deaths from the same source, to simulate reported cases and deaths.

Just like Bosse *et al.*[3] and Farrow *et al.*[7], this study struggled to retain a large number of participants. Focused public outreach efforts such as creating a dedicated website, announcing an official tournament, providing a public leaderboard, sending weekly emails with details on past performance and weekly announcements on Twitter, did noticeably increase participation compared to the previous study in Germany and Poland. Nevertheless, retaining participants beyond the initial recruitment proved challenging, and most forecasters only submitted a single forecast. McAndrew *et al.*[5] had a higher number of participants, suggesting that making use of existing forecasting platforms that have access to a large existing user base and greater resources may be helpful in recruiting a larger number of participants, though these platforms lack the flexibility and software tooling to run a novel study of this kind in real-time as things stand.

The study period was marked by an increase in both case and death numbers. Case numbers rose quickly compared to the pre-study period, peaking on July 17, 2021, followed by a trough and a subsequent further increase. Forecasts displayed a pattern where forecasters tended to underpredict while case numbers were rising, and overpredict while case numbers were falling, particularly following a peak. Similar patterns have been observed previously in other short-term forecasts of COVID-19 (see e.g. 3,9,11).

Death numbers during the study period were increasing more slowly than during the previous peak in January 2021, coinciding with the beginning of vaccination efforts and a growing immunity in the population[28]. The peak in case numbers in July 2021 was not followed by a subsequent peak in death numbers (but rather a steady incline over several months), suggesting some decoupling of case and death numbers such as would be expected from effects of immunity that are stronger in preventing severe disease than any symptoms. Forecasters tended to overpredict death numbers in the beginning, while underpredicting them in the end, expecting death numbers to fall after the peak in cases. The study period coincides with

the rise of the Delta variant in the UK[27,28], as well as the 2021 European Football Championship, which likely shifted the age distribution towards younger cases[29].

In line with results from previous work[3,11], we found almost all forecasts for cases to be underdispersed (i.e., too narrow/overconfident). Empirical coverage for death forecasts was higher than the corresponding coverage for cases for all forecasting approaches except the R_t forecasts.

For forecasts of cases two weeks ahead, performance of the human judgement forecasts was better than the European Forecast Hub ensemble in terms of WIS on the log scale, and worse in terms of WIS on the natural scale. This was linked to a tendency of the Hub ensemble to make lower case predictions, which led to lower overprediction penalties on the natural scale, but noticeably higher underprediction penalties on the log scale. For forecasts of deaths two weeks ahead, direct human forecasts and the combined crowd ensemble performed better than the Hub ensemble on the log scale. On the natural scale, the combined crowd ensemble performed worse than the Hub ensemble, while the direct crowd forecasts still performed better. R_t forecasts for deaths performed noticeably worse than all other approaches both on the log and on the natural scale.

In their original study, conducted in Germany and Poland, Bosse *et al.*[3] found that humans outperformed an ensemble of computational models when predicting cases, but not when predicting deaths. They hypothesised that computational models might have an advantage over human forecasters when predicting deaths, benefiting from the ability to model the delays and epidemiological relationships between different leading and lagged indicators. McAndrew *et al.*[5] similarly found in their study that humans performed comparably to an ensemble of computational models for cases, but not for predictions of deaths of COVID-19. Results in our study do not directly support this pattern, but given the low number of observations also do not provide strong evidence against it. In this study, the combined crowd ensemble performed better than the Hub ensemble on both cases and deaths on the log scale, and worse on the natural scale. Direct forecasts, which would be most comparable to the forecasts in Bosse *et al.*[3], performed worse than the Hub ensemble on cases and better on deaths. During the study period, the case fatality ratio (CFR) likely changed quite quickly compared to the pre-study period. On the one hand, the rise of the Delta variant in the UK, which was first detected in the UK in March 2021 was estimated to have a higher CFR than previous variants[27,31] (although Perez-Guzman *et al.*[28] estimated it to be lower than that of the Alpha variant). On the other hand, the ongoing COVID-19 vaccination and growing natural immunity in the population had decreasing effects on the CFR. In addition, the age distribution of cases changed (hence modifying the overall CFR) throughout study period in Summer 2021, in parts related to the European Football Championship[29]. Overall, the CFR was lower than during previous peaks of COVID-19[28]. One possible hypothesis for the relatively good performance of human forecasts for deaths compared to previous studies might be that some models submitted to the Forecast Hub may have been more negatively affected by the changes in CFR during the study period than human forecasters or have been slower to update. The present study only saw a steady increase in death numbers, which one could argue is relatively easy to predict, making it difficult to compare forecast performance with performance in other settings. A confounding factor, when comparing results from this study and the one in Germany and Poland directly, is that we used a median ensemble to combine individual forecasts here, while the earlier study used a mean ensemble.

Importantly, in this study our combined crowd ensemble ("epiforecasts-EpiExpert") contributed to the European Forecast Hub ensemble. This is in contrast to the study by Bosse *et al.*[3], where they compared crowd forecasts against a hypothetical ensemble excluding the crowd forecasts. In the original study, including the crowd forecasts improved the Hub ensemble on average (however, the overall number of models included in the German and Polish Hub ensemble was smaller than the number of models in the European Forecast Hub ensemble). In our study, comparisons between our crowd ensembles and the Forecast Hub ensemble are therefore confounded by the fact the combined crowd ensemble was included in the Forecast Hub ensemble, possibly leading us to underestimate differences between the two.

This study explored a novel method of forecasting infectious diseases that combines a human forecast of the estimated effective reproduction number R_t with epidemiological modelling to map the R_t forecast to a forecast of cases and deaths. One appeal of this approach is that the forecaster can directly forecast the generative process and how they believe it is affected by interventions and changes in behaviour. Computational modelling then takes care of dealing with details such as reporting delays, generation intervals, day of the week periodicity, and the relationship between different indicators. This could help reduce cognitive load, and make it easier to synthesise various sources in information into a single forecast, at least for forecasters who have an intuitive understanding of R_t. Though we note all of these modelling steps and the construction of the model itself requires the human constructing the model to make assumptions. Anecdotally, forecasters familiar to the authors reported high satisfaction with the forecasting experience. One important limitation of the approach is that R_t values were estimated based on reported numbers of cases. This is susceptible to changes in testing and reporting and estimated R_t values may not accurately reflect the true underlynig infectious disease dynamics. In our study, R_t forecasts of cases were comparable to direct forecasts, with a tendency for good forecasters to improve when using the R_t method and worse forecasters to deteriorate even more. Sample sizes, however, were very low. Given that forecasters could simulate cases in the app, it is also possible that forecasters were in fact directly forecasting cases. R_t forecasts of deaths (which forecasters could not see in the app) were noticeably worse than direct forecasts of deaths. The computational model underlying our R_t forecasts of deaths estimated a constant CFR and delay distribution using the last 4 weeks of data, therefore updating relatively slowly to

new circumstances and the CFR was assumed to be constant over the four week forecast horizon. However, as mentioned before, the CFR likely evolved during the study period. Forecasters had no way of inspecting the death forecast implied by their R_t forecast, likely impacting predictive performance. They also had no way to adjust the CFR manually, likely impacting forecast accuracy. Allowing human forecasters to see their implied death forecasts, as well as giving them the ability to adjust the CFR and other model parameters would have increased complexity of the interface, but would have solved issues with the assumptions of the underlying model. Alternatively, a more complex model could have been used which allowed for time-varying CFR estimates and forecast these changes over the forecast horizon though this approach may still have struggled to cope with the rapid changes observed during the study period. Another important limitation is that we didn't have full sample trajectories of the R_t-values predicted by forecasters. Rather, trajectories had to be constructed based on the distributions provided for the different forecast horizons, which likely negatively affected forecasts. One potential way to disentangle the effect of the convolution model from the R_t forecasts would have been to use the human forecasts for cases as an input to the second computational model, which could then have simulated deaths. Future work could expose forecasters to different combinations of these options with the aim of separating effects of the user interface from ones related to the structure of the underlying computational model.

Combining forecasts from "experts" and "non-experts" led to better performance for forecasts two weeks ahead for cases as well as deaths, and both on the log scale and on the natural scale. Combining direct forecasts and R_t forecasts led to better performance on the log scale, but not on the natural scale. This suggests that combining different forecasts can be beneficial in many instances, although there may be differences in terms of WIS on the log and the natural scale. In particular, WIS values on the natural scale may be more susceptible to models that would tend to overshoot and miss the peak, while WIS on the log scale may be more affected by models that underpredict and miss upswings[25].

Past studies of expert forecasts of COVID-19[6] had found predictions from experts to outperform those of non-experts. In our study, an ensemble of self-reported "experts" outperformed an ensemble of "non-experts" when forecasting cases two weeks ahead, both on the log scale and on the natural scale. When forecasting deaths two weeks ahead, "experts" performed worse than "non-experts" on the log scale, but better on the natural scale. Forecasts for "experts" tended to be better calibrated than non-experts. However results should be taken with care considering relatively low sample sizes (median of 11 "non-experts" for cases and 10 for deaths, median of 8 "experts" for cases and deaths) and given that expert status was self-reported. Furthermore, we only asked for professional involvement in a field related to infectious disease modelling, not specifically for familiarity with modelling of COVID-19 in the UK, and only offered participants a binary choice. However, as we used ad-hoc recruitment in our networks many of these self-identified experts are likely to be infectious disease modellers.

It is plausible to hypothesise that the default baseline shown to forecasters in the app may influence their predictions. One could also interpret the Rt-forecast as a way of showing a different baseline forecast to the forecaster compared to the direct forecast. In our study, the default was a naive forecast with the median equal to the last value and uncertainty equal to the standard deviation of the last four changes in weekly log values. Bosse *et al.*[3] did not find conclusive evidence to that effect, but also did not analyse the question in detail. We suggest further research be done into potential priming effects that a default forecast can have on users.

Overall, results of our study should be taken with caution due to several important limitations. Firstly, our study was restricted to one location and to a relatively short period of thirteen weeks. Secondly, there were many confounding factors that likely influence results. These include the fact that different participants made forecasts at different points in time (with the median forecaster only submitting a single forecast) and that subgroups of interest (e.g. "experts", or R_t forecasts) had different numbers of forecasters. In most instances, differences in scores between forecast approaches were small compared to the variance of scores within a single approach. In addition, there were many researcher degrees of freedom that could influence findings, for example how individual forecasts were combined to create an ensemble. Results were influenced by choices made during the evaluation with, for example, some conclusions depending on forecast horizon and the transformation used prior to scoring. Highlighting this, prizes to the human forecasters were paid out based on the combined WIS on the log scale across all horizons and forecast targets. Had we chosen to instead measure WIS on the natural scale, or to forecast only cases and continue to score on the log scale, rankings and payouts would have been different.

Conclusions

The results of our study are broadly consistent with previous studies on human judgement forecasting of COVID-19 and suggest that human crowd ensembles and an ensemble of computational models are able to produce forecasts of similar quality. One interpretation of these findings is that a mixed crowd of human forecaster can produce a viable alternative or complement to an ensemble of mathematical models created by experts. An altnerative interpretation is that an ensemble of automated models can produce forecasts over the course of several years that are on par with that of an engaged crowd of human forecasters. This study, and all previous studies, comparing human judgement forecasts and computational models only ran over short periods of time and the majority of them struggled with recruitment and upkeep. Meanwhile,

COVID-19 Forecast Hubs have attracted continuous submissions for almost three years and were able to consistently provide forecasts of comparable quality.

Our findings do not suggest that humans are necessarily at a general disadvantage compared to computational models at predicting reported deaths, but evidence in both directions is limited and this is made particularly complex as our study took place during a period of time when CFR estimates were changing rapidly. Despite evaluations being public, it remains a challenge to properly incentivise contributors to Forecast Hubs to regularly update their forecasting methodology in order to maximise utility, predictive performance, or both. Combining human judgement and epidemiological modelling by mapping R_t forecasts to case and death numbers has not yielded competitive forecasts for deaths in this study. However, we only presented a prototype of a forecasting approach, which, while having appealing properties, proved challenging to implement. Subsequent iterations and improvements could likely achieve better results. More research is required to obtain a better understanding of the role of subject matter expertise in infectious disease forecasting. Similarly, it would be interesting to explore the effects on predictive accuracy of providing forecasters with additional qualitative real-time information such as detailed descriptive reports that enhance the forecasters' understanding of the overall context beyond the numerical data that was visible in our application. Our results underline that it is difficult to evaluate forecast performance devoid of context that helps inform what a good or a bad forecast is. Different ways to look at the data let different forecasts appear better or worse. Forecast evaluation therefore either needs to be clearly informed by the needs of forecast consumers to determine what a good forecast is, or it needs a broad array of perspectives to provide a wholistic picture as we have attempted to present in this work. Furthermore, evaluating forecasts post-hoc leaves the researchers with many degrees of freedom to make decisions that affect which models look good and there is a risk of allowing for motivated reasoning. More emphasis should be put on measures that prevent this, e.g. by establishing common standards for evaluations, pre-registering studies, and making it a norm to display a variety of standard metrics.

Data availability

All data and code are available publicy under a MIT license under https://github.com/epiforecasts/uk-crowd-forecasting-challenge and https://doi.org/10.5281/zenodo.7897257. The data has been published separately here: https://doi.org/10.5281/zenodo.7897289. Supplementary Information[12] to this manuscript is available at https://doi.org/10.5281/zenodo.7897513.

Author contributions

NIB contributed to the conceptualization, data curation, formal analysis, investigation, methodology, software development, validation, visualization, and original draft preparation of the manuscript, as well as its review and editing. SA contributed to the conceptualization, data curation, investigation, methodology, software development, supervision, and review and editing of the manuscript. JB contributed to the supervision, review, and editing of the manuscript. EvL contributed to the conceptualization, supervision, and review and editing of the manuscript. AC contributed to the conceptualization, supervision, and review and editing of the manuscript. SF contributed to the conceptualization, funding acquisition, project administration, supervision, and review and editing of the manuscript. All authors have read and approved the final version of the manuscript.

Acknowledgements

We thank all forecasters for their participation and want to congratulate the three winners of the forecasting challenge: Russell Bradshaw, Sebastian Funk (an author of this study), and Akira Endo. All winners donated their prizes. We also thank Daniel J. McDonald and Hongru Du for their kind and thoughtful reviews which have helped improve the manuscript substantially.

References

1. Cramer E, Ray EL, Lopez VK, *et al.*: **Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the US.** *medRxiv.* 2021.02.03.21250974, 2021.
Reference Source

2. Venkatramanan S, Cambeiro J, Liptay T, *et al.*: **Utility of human judgment ensembles during times of pandemic uncertainty: A case study during the COVID-19 Omicron BA.1 wave in the USA**. 2022.10.12.22280997, 2022.
Reference Source

3. Bosse NI, Abbott S, Bracher J, *et al.*: **Comparing human and model-based forecasts of COVID-19 in Germany and Poland.** *PLoS Comput Biol.* 2022; **18**(9): e1010405.
PubMed Abstract | Publisher Full Text | Free Full Text

4. McAndrew T, Reich NG: **An expert judgment model to predict early stages of the COVID-19 pandemic in the United States.** *PLoS Comput Biol.* 2022; **18**(9): e1010485.
PubMed Abstract | Publisher Full Text | Free Full Text

5. McAndrew T, Codi A, Cambeiro J, *et al.*: **Chimeric forecasting: combining probabilistic predictions from computational models and human judgment.** *BMC Infect Dis.* 2022; **22**(1): 833.
PubMed Abstract | Publisher Full Text | Free Full Text

6. Recchia G, Freeman ALJ, Spiegelhalter D: **How well did experts and laypeople forecast the size of the COVID-19 pandemic?** *PLoS One.* 2021; **16**(5): e0250935.
PubMed Abstract | Publisher Full Text | Free Full Text

7. Farrow DC, Brooks LC, Hyun S, *et al.*: **A human judgment approach to epidemiological forecasting.** *PLoS Comput Biol.* 2017; **13**(3): e1005248.
PubMed Abstract | Publisher Full Text | Free Full Text

8. Swallow B, Birrell P, Blake J, *et al.*: **Challenges in estimation, uncertainty quantification and elicitation for pandemic modelling.** *Epidemics.* 2022; **38**: 100547.
PubMed Abstract | Publisher Full Text | Free Full Text

9. Bracher J, Wolffram D, Deuschel JK, *et al.*: **A pre-registered short-term forecasting study of COVID-19 in Germany and Poland during the second wave.** *Nat Commun.* 2021; **12**(1): 5173.
PubMed Abstract | Publisher Full Text | Free Full Text

10. Bracher J, Wolffram D, Deuschel J, *et al.*: **National and subnational short-term forecasting of COVID-19 in Germany and Poland during early 2021.** *Commun Med (Lond).* 2022; **2**(1): 136.
PubMed Abstract | Publisher Full Text | Free Full Text

11. Sherratt K, Gruson H, Grah R, *et al.*: **Predictive performance of multi-model ensemble forecasts of COVID-19 across European nations.** *eLife.* 2023; **12**: e81916.
PubMed Abstract | Publisher Full Text | Free Full Text

12. Bosse N, Abbott S, Bracher J, *et al.*: **Supplementary Information - Human Judgement forecasting of COVID-19 in the UK**. 2023.
http://www.doi.org/10.5281/zenodo.7897513

13. R Core Team: **R: A Language and Environment for Statistical Computing**. R Foundation for Statistical Computing, Vienna, Austria, 2022.
Reference Source

14. Chang W, Cheng J, Allaire JJ, *et al.*: **shiny: Web Application Framework for R**. R package version 1.6.0, 2021.
Reference Source

15. Bosse N, Abbott S, Funk S: **epiforecasts/crowdforecastr: beta release**. 2021.
Publisher Full Text

16. Mathieu E, Ritchie H, Rodés-Guirao L, *et al.*: **Coronavirus pandemic (covid-19). Our World in Data**. 2020.
Reference Source

17. Abbott S, Hellewell J, Sherratt K, *et al.*: **EpiNow2: Estimate Real-Time Case Counts and Time-Varying Epidemiological Parameters**. 2020.
Reference Source

18. Fraser C: **Estimating Individual and Household Reproduction Numbers in an Emerging Epidemic.** *PLoS One.* 2007; **2**(8): e758.
PubMed Abstract | Publisher Full Text | Free Full Text

19. Abbott S, CMMID COVID-19 Working Group, Kucharski AJ, *et al.*: **Estimating the increase in reproduction number associated with the Delta variant using local area dynamics in England**. 2021.11.30.21267056; 2021.
Reference Source

20. Abbott S, Hellewell J, Thompson RN, *et al.*: **Estimating the time-varying reproduction number of SARS-CoV-2 using national and subnational case counts [version 1; peer review: awaiting peer review].** *Wellcome Open Res.* 2020; **5**: 112.
Publisher Full Text

21. Sherratt K, Abbott S, Meakin SR, *et al.*: **CMMID Covid-19 working Group, Mark Jit and Sebastian Funk. Exploring surveillance data biases when estimating the reproduction number: With insights into subpopulation transmission of Covid-19 in England**. 2020.10.18.20214585, 2021.
Reference Source

22. Ray EL, Brooks LC, Bien J, *et al.*: **Comparing trained and untrained probabilistic ensemble forecasts of COVID-19 cases and deaths in the United States.** *Int J Forecast.* 2023; **39**(3): 1366–1383.
PubMed Abstract | Publisher Full Text | Free Full Text

23. Bracher J, Ray EL, Gneiting T, *et al.*: **Evaluating epidemic forecasts in an interval format.** *PLoS Comput Biol.* 2021; **17**(2): e1008618.
PubMed Abstract | Publisher Full Text | Free Full Text

24. Gneiting T, Balabdaoui F, Raftery AE: **Probabilistic forecasts, calibration and sharpness.** *J R Statist Soc B.* 2007; **69**(Part 2): 243–268.
Reference Source

25. Bosse NI, Abbott S, Cori A, *et al.*: **Scoring epidemiological forecasts on transformed scales**. 2023.
Reference Source

26. Bosse NI, Gruson H, Cori A, *et al.*: **Evaluating Forecasts with scoringutils in R**. 2022.
Reference Source

27. Bast E, Tang F, Dahn J, *et al.*: **Increased risk of hospitalisation and death with the delta variant in the USA.** *Lancet Infect Dis.* 2021; **21**(12): 1629–1630.
PubMed Abstract | Publisher Full Text | Free Full Text

28. Perez-Guzman PN, Knock E, Imai N, *et al.*: **Epidemiological drivers of transmissibility and severity of SARS-CoV-2 in England.** *Nat Commun.* ISSN 2041-1723, 2023; **14**(1): 4279.
PubMed Abstract | Publisher Full Text | Free Full Text

29. Dehning J, Mohr SB, Contreras S, *et al.*: **Impact of the Euro 2020 championship on the spread of COVID-19.** *Nat Commun.* 2023; **14**(1): 122.
PubMed Abstract | Publisher Full Text | Free Full Text

30. Kendall M, Tsallis D, Wymant C, *et al.*: **Epidemiological impacts of the NHS COVID-19 app in England and Wales throughout its first year.** *Nat Commun.* ISSN 2041-1723, 2023; **14**(1): 858.
PubMed Abstract | Publisher Full Text | Free Full Text

31. Twohig KA, Nyberg T, Zaidi A, *et al.*: **Hospital admission and emergency care attendance risk for SARS-CoV-2 delta (B.1.617.2) compared with alpha (B.1.1.7) variants of concern: A cohort study.** *Lancet Infect Dis.* ISSN 1473-3099, 1474-4457, 2022; **22**(1): 35–42.
PubMed Abstract | Publisher Full Text | Free Full Text

Supplementary information

Weighted interval score

The weighted interval score (smaller values are better) is a proper scoring rule for quantile forecasts. It converges to the continuous ranked probability score (which itself is a generalisation of the absolute error to probabilistic forecasts) for an increasing number of intervals. The score can be decomposed into a dispersion (uncertainty) component and penalties for over- and underprediction. For a single interval, the score is computed as

$$IS_\alpha(F,y) = (u-l) + \frac{2}{\alpha}\cdot(l-y)\cdot 1(y \le l) + \frac{2}{\alpha}\cdot(y-u)\cdot 1(y \ge u),$$

where 1() is the indicator function, y is the true value, and l and u are the $\frac{\alpha}{2}$ and $1-\frac{\alpha}{2}$ quantiles of the predictive distribution F, i.e., the lower and upper bound of a single prediction interval. For a set of K prediction intervals and the median m, the score is computed as a weighted sum,

$$WIS = \frac{1}{K+0.5}\cdot\left(w_0\cdot|y-m| + \sum_{k=1}^{K} w_k \cdot IS_\alpha(F,y)\right),$$

where w_k is a weight for every interval. Usually, $w_k = \frac{\alpha_k}{2}$ and $w_0 = 0.5$.

Renewal equation model

The model was initialised prior to the first observed data point by assuming constant exponential growth for the mean of assumed delays from infection to case report.

$$I_t = I_0 \exp(rt) \tag{1}$$
$$I_0 \sim \mathscr{LN}(\log I_{obs}, 0.2) \tag{2}$$
$$r \sim \mathscr{LN}(r_{obs}, 0.2) \tag{3}$$

Where I_{obs} and r_{obs} are estimated from the first week of observed data. For the time window of the observed data infections were then modelled by weighting previous infections by the generation time and scaling by the instantaneous reproduction number. These infections were then convolved to cases by date (O_t) and cases by date of report (D_t) using log-normal delay distributions. This model can be defined mathematically as follows,

$$\log R_t = \log R_{t-1} + \mathrm{GP}_t \tag{4}$$
$$I_t = R_t \sum_{\tau=1}^{15} w(\tau|\mu_w,\sigma_w) I_{t-\tau} \tag{5}$$
$$O_t = \sum_{\tau=0}^{15} \xi_O(\tau|\mu_{\xi_O},\sigma_{\xi_O}) I_{t-\tau} \tag{6}$$
$$D_t = \alpha \sum_{\tau=0}^{15} \xi_D(\tau|\mu_{\xi_D},\sigma_{\xi_D}) O_{t-\tau} \tag{7}$$
$$C_t \sim \mathrm{NB}\left(\omega_{(t \mod 7)} D_t, \phi\right) \tag{8}$$

Where,

$$w \sim \mathscr{G}(\mu_w,\sigma_w) \tag{9}$$
$$\xi_O \sim \mathscr{LN}(\mu_{\xi_O},\sigma_{\xi_O}) \tag{10}$$
$$\xi_D \sim \mathscr{LN}(\mu_{\xi_D},\sigma_{\xi_D}) \tag{11}$$

This model used the following priors for cases,

$$R_0 \sim \mathscr{LN}(0.079, 0.18) \tag{12}$$
$$\mu_w \sim \mathscr{N}(3.6, 0.7) \tag{13}$$
$$\sigma_w \sim \mathscr{N}(3.1, 0.8) \tag{14}$$
$$\mu_{\xi_O} \sim \mathscr{N}(1.62, 0.064) \tag{15}$$
$$\sigma_{\xi_O} \sim \mathscr{N}(0.418, 0.069) \tag{16}$$
$$\mu_{\xi_D} \sim \mathscr{N}(0.614, 0.066) \tag{17}$$
$$\sigma_{\xi_D} \sim \mathscr{N}(1.51, 0.048) \tag{18}$$
$$\alpha \sim \mathscr{N}(0.25, 0.05) \tag{19}$$
$$\frac{\omega}{7} \sim \mathrm{Dirichlet}(1,1,1,1,1,1,1) \tag{20}$$
$$\phi \sim \frac{1}{\sqrt{\mathscr{N}(0,1)}} \tag{21}$$

and updated the reporting process as follows when forecasting deaths,

$$\mu_{\xi_D} \sim \mathscr{N}(2.29, 0.076) \tag{22}$$
$$\sigma_{\xi_D} \sim \mathscr{N}(0.76, 0.055) \tag{23}$$
$$\alpha \sim \mathscr{N}(0.005, 0.0025) \tag{24}$$

α, μ, σ, and ϕ were truncated to be greater than 0 and with ξ, and w normalised to sum to 1.

The prior for the generation time was sourced from [5] but refit using a log-normal incubation period with a mean of 5.2 days (SD 1.1) and SD of 1.52 days (SD 1.1) with this incubation period also being used as a prior [6] for ξ_O. This resulted in a gamma-distributed generation time with mean 3.6 days (standard deviation (SD) 0.7), and SD of 3.1 days (SD 0.8) for all estimates. We estimated the delay between symptom onset and case report or death required to convolve latent infections to observations by fitting an integer adjusted log-normal distribution to 10 subsampled bootstraps of a public linelist for cases in Germany from April 2020 to June 2020 with each bootstrap using 1% or 1769 samples of the available data [11, 2] and combining the posteriors for the mean and standard deviation of the log-normal distribution [1, 4, 9, 10].

GP_t is an approximate Hilbert space Gaussian process as defined in [7] using a Matern 3/2 kernel using a boundary factor of 1.5 and 17 basis functions (20% of the number of days used in fitting). The length scale of the Gaussian process was given a log-normal prior with a mean of 21 days, and a standard deviation of 7 days truncated to be greater than 3 days and less than 60 days. The magnitude of the Gaussian process was assumed to be normally distributed centred at 0 with a standard deviation of 0.1. From the forecast time horizon (T) and onwards the last value of the Gaussian process was used (hence R_t was assumed to be fixed) and latent infections were adjusted to account for the proportion of the population that was susceptible to infection as follows,

$$I_t = (N - I^c_{t-1})\left(1 - \exp\left(\frac{-I'_t}{N - I^c_T}\right)\right), \tag{25}$$

where $I_t^c = \sum_{s<t} I_s$ are cumulative infections by $t-1$ and I'_t are the unadjusted infections defined above. This adjustment is based on that implemented in the `epidemia` R package [8, 3].

Convolution model The convolution model shares the same observation model as the renewal model but rather than assuming that an observation is predicted by itself using the renewal equation instead assumes that it is predicted entirely by another observation after some parametric delay. It can be defined mathematically as follows,

$$D_t \sim \text{NB}\left(\omega_{(t \mod 7)}\alpha \sum_{\tau=0}^{30} \xi(\tau|\mu,\sigma) C_{t-\tau}, \phi\right) \tag{26}$$

with the following priors,

$$\frac{\omega}{7} \sim \text{Dirichlet}(1,1,1,1,1,1,1) \tag{27}$$
$$\alpha \sim \mathcal{N}(0.01, 0.02) \tag{28}$$
$$\xi \sim \mathcal{LN}(\mu,\sigma) \tag{29}$$
$$\mu \sim \mathcal{N}(2.5, 0.5) \tag{30}$$
$$\sigma \sim \mathcal{N}(0.47, 0.2) \tag{31}$$
$$\phi \sim \frac{1}{\sqrt{\mathcal{N}(0,1)}} \tag{32}$$

with α, μ, σ, and ϕ truncated to be greater than 0 and with ξ normalised such that $\sum_{\tau=0}^{30} \xi(\tau|\mu,\sigma) = 1$.

Model fitting

Both models were implemented using the `EpiNow2` R package (version 1.3.3) [1]. Each forecast target was fitted independently for each model using Markov-chain Monte Carlo (MCMC) in stan [10]. A minimum of 4 chains were used with a warmup of 250 samples for the renewal equation-based model and 1000 samples for the convolution model. 2000 samples total post warmup were used for the renewal equation model and 4000 samples for the convolution model. Different settings were chosen for each model to optimise compute time contingent on convergence. Convergence was assessed using the R hat diagnostic [10]. For the convolution model forecast the case forecast from the renewal equation model was used in place of observed cases beyond the forecast horizon using 1000 posterior samples. 12 weeks of data was used for both models though only 3 weeks of data were included in the likelihood for the convolution model.

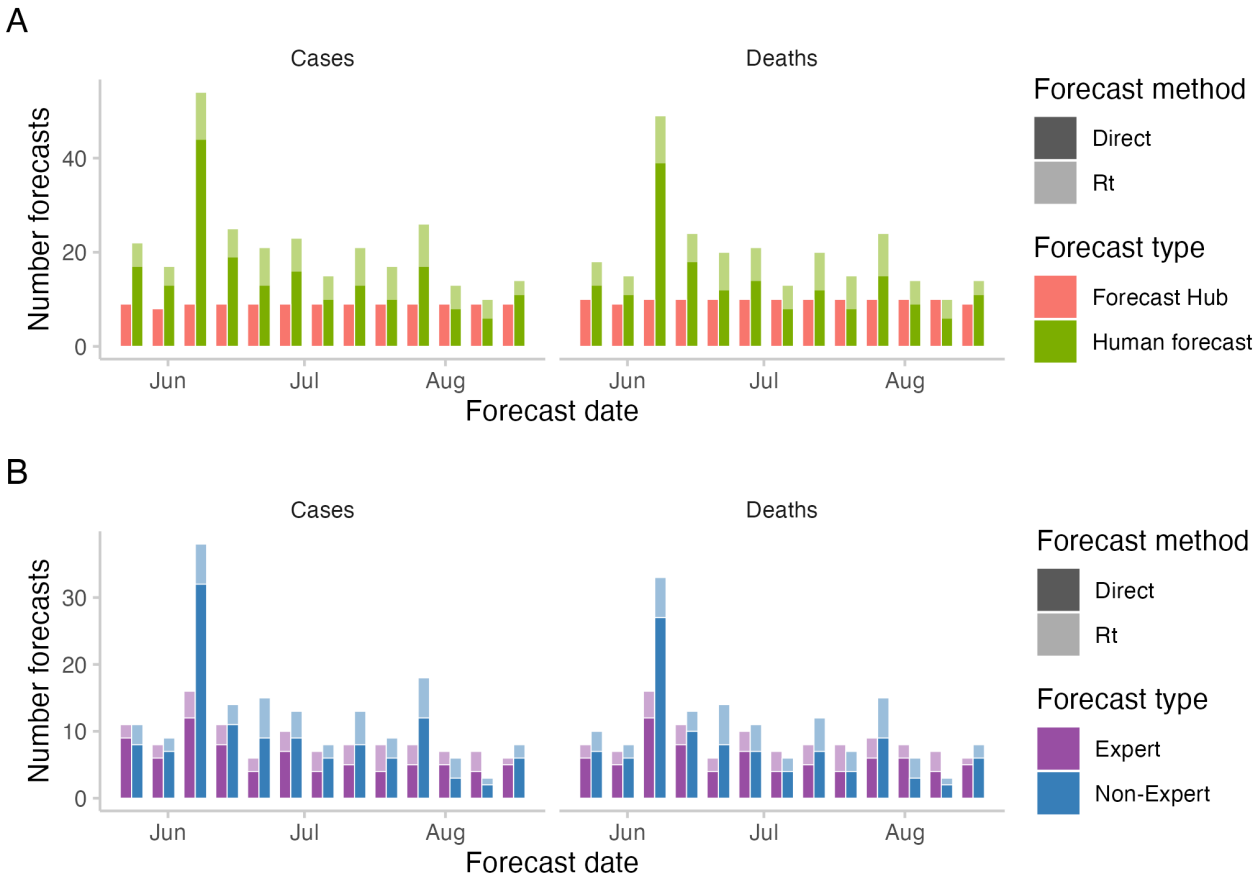

Figure SI.1. Number of forecasts across the study period. A: number of forecasts included in the Hub ensemble and the combined crowd ensemble. B: number of forecasts by "experts" and "non-experts". Expert status was determined based on the participant's answer to the question whether they "worked in infectious disease modelling or had professional experience in any related field".

Table SI.1. Performance for four-week-ahead forecasts. Values have been cut to three significant digits and rounded

		WIS - natural			WIS - log scale				
Model	Target	abs.	rel.	sd	abs.	rel.	sd	Coverage 50%	Coverage 90%
EuroCOVIDhub-ensemble	Cases	81.5k	1	74.8k	0.45	1	0.3	0.23	0.62
crowd-ensemble	Cases	98.4k	1.21	115k	0.45	1.02	0.35	0.23	0.62
crowd-direct	Cases	101k	1.25	128k	0.47	1.05	0.41	0.31	0.62
crowd-rt	Cases	106k	1.3	118k	0.47	1.06	0.33	0.23	0.46
EuroCOVIDhub-ensemble	Deaths	85	1	60.3	0.2	1	0.08	0.77	0.92
crowd-ensemble	Deaths	98.2	1.16	103	0.19	0.95	0.13	0.54	0.77
crowd-direct	Deaths	88.1	1.04	80.8	0.22	1.08	0.14	0.38	0.77
crowd-rt	Deaths	154	1.82	110	0.31	1.51	0.17	0.15	0.46

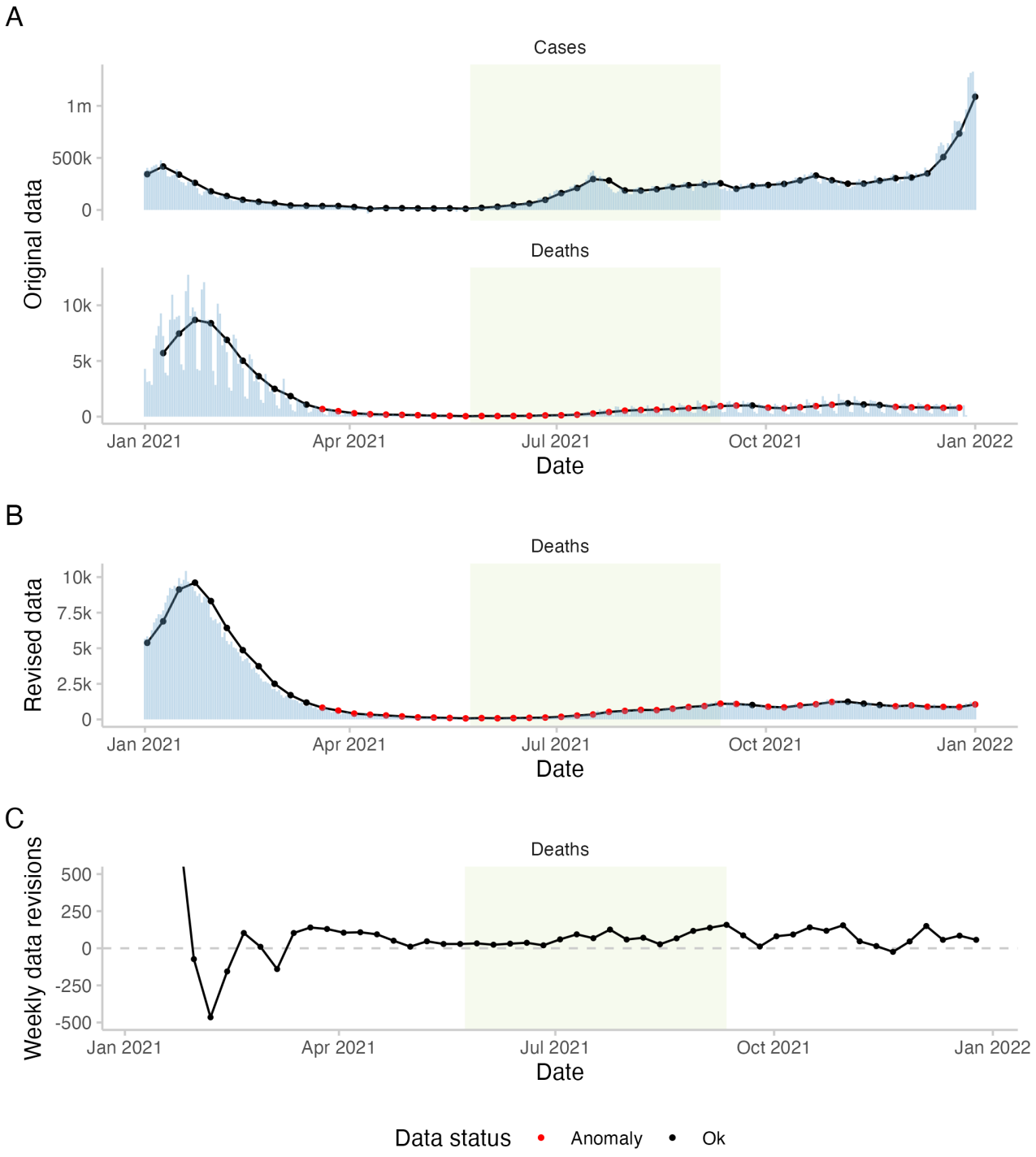

Figure SI.2. **Observed cases and deaths of COVID-19 in the UK. A: Observed daily (bars) and weekly (black lines and points) numbers of cases and deaths as available through the European Forecast Hub when the study concluded in 2021. Daily numbers were multiplied by seven in order to appear on the same scale as weekly numbers. Red dots represent days for which the original data and the revised data disagreed by more than five percent. B: Revised data available as of February 14 2023. In August, Johns Hopkins University that provided the data switched the data stream for their death forecasts to reflect the number of death certificates that mentioned COVID-19 rather than the number of people who died within 28 days of a positive test. C: Difference between the original and revised weekly death numbers.**

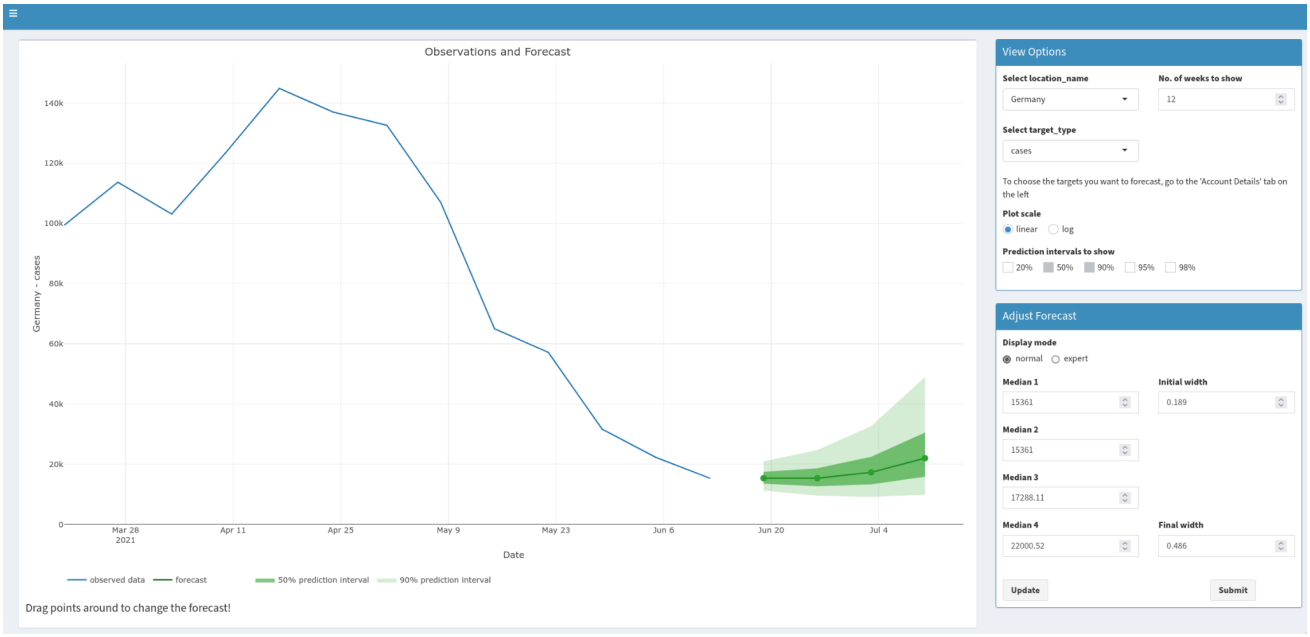

Figure SI.3. **Screenshot of the direct forecasting interface.**

Figure SI.4. **Screenshot of the R_t forecasting interface.**

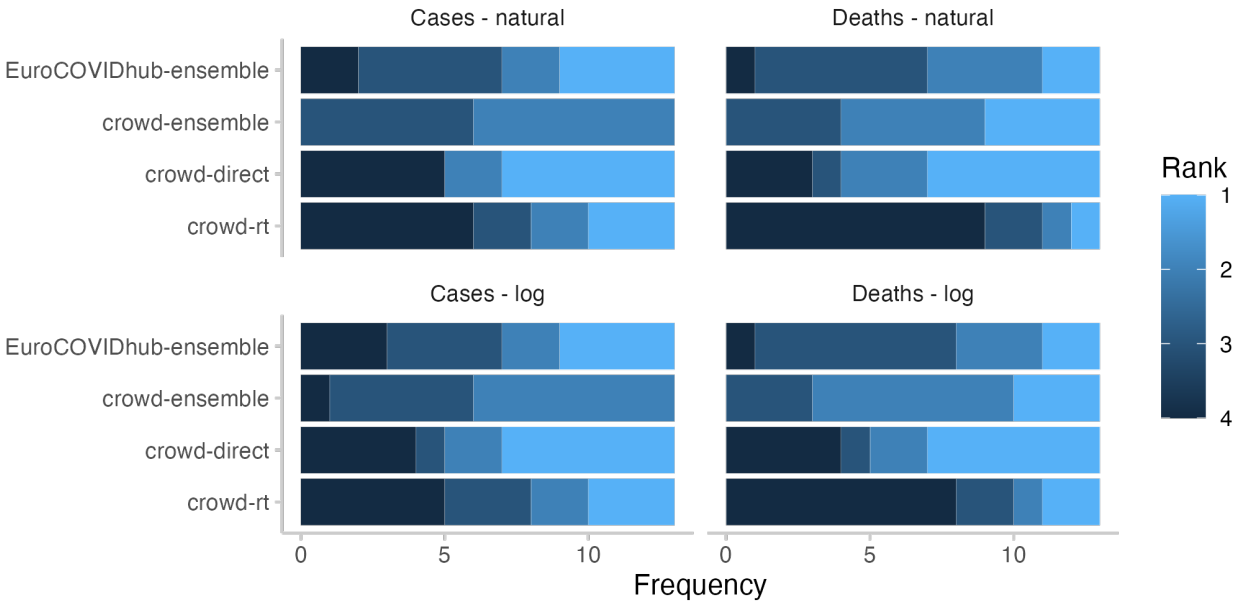

Figure SI.5. Ranks for all forecasting approaches for four week ahead forecasts. Colours indicate how often (out of 13 forecasts) a given approach got 1st, 2nd, 3rd, or 4th rank.

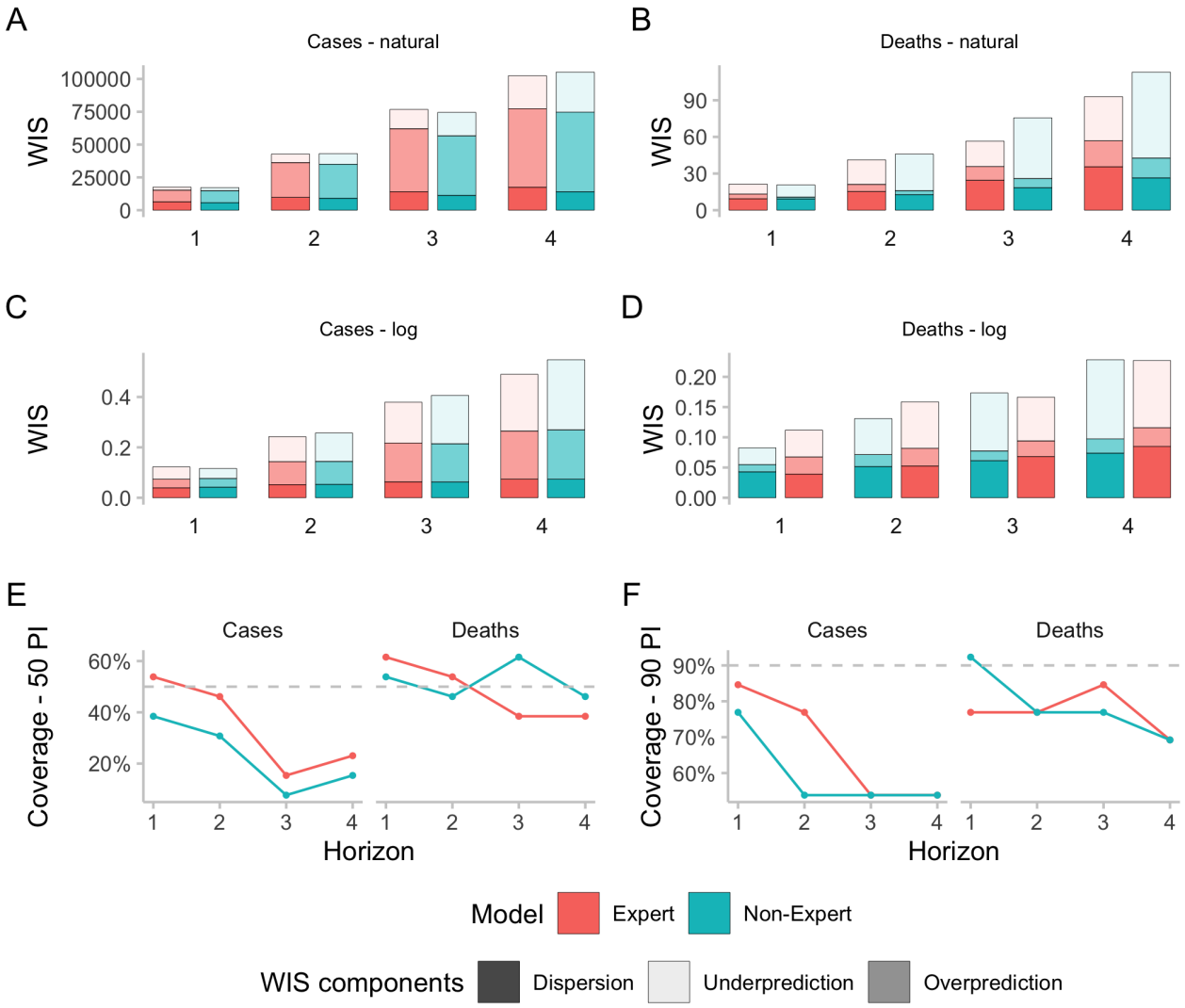

Figure SI.6. **Predictive performance of self-reported "experts" and "non-experts" across forecast horizons. Forecasts from "experts" and "non-experts" were combined to two separate median ensembles, including both direct and R_t forecasts. A-D: WIS stratified by forecast horizon for cases and deaths on the natural and log scale. E, F: Empirical coverage of the 50% and 90% prediction intervals stratified by forecast horizon and target type.**

Table SI.2. **Performance for two-week-ahead forecasts of experts and non-experts. Values have been cut to three significant digits and rounded.**

		WIS - natural			WIS - log scale				
Model	Target	abs.	rel.	sd	abs.	rel.	sd	Coverage 50%	Coverage 90%
crowd-ensemble	Cases	40.1k	1	69.4k	0.22	1	0.25	0.38	0.69
Expert	Cases	42.7k	1.06	74.9k	0.24	1.08	0.28	0.46	0.77
Non-Expert	Cases	43.1k	1.07	67k	0.26	1.14	0.25	0.31	0.54
crowd-ensemble	Deaths	40.2	1	41.5	0.12	1	0.07	0.54	0.77
Expert	Deaths	41.2	1.03	41.8	0.16	1.29	0.15	0.54	0.77
Non-Expert	Deaths	45.9	1.14	56.8	0.13	1.06	0.08	0.46	0.77

Table SI.3. Performance for four-week-ahead forecasts of experts and non-experts. Values have been cut to three significant digits and rounded.

		WIS - natural			WIS - log scale				
Model	Target	abs.	rel.	sd	abs.	rel.	sd	Coverage 50%	Coverage 90%
Expert	Cases	102k	1.04	121k	0.49	1.08	0.4	0.23	0.54
Non-Expert	Cases	105k	1.07	110k	0.55	1.21	0.4	0.15	0.54
crowd-ensemble	Cases	98.4k	1	115k	0.45	1	0.35	0.23	0.62
Expert	Deaths	93	0.95	81.2	0.23	1.17	0.14	0.38	0.69
Non-Expert	Deaths	113	1.15	122	0.23	1.18	0.18	0.46	0.69
crowd-ensemble	Deaths	98.2	1	103	0.19	1	0.13	0.54	0.77

References

[1] Sam Abbott, Joel Hellewell, Katharine Sherratt, Katelyn Gostic, Joe Hickson, Hamada S. Badr, Michael DeWitt, Robin Thompson, EpiForecasts, and Sebastian Funk. *EpiNow2: Estimate Real-Time Case Counts and Time-Varying Epidemiological Parameters*, 2020.

[2] Sam Abbott, Katharine Sherratt, Jonnie Bevan, Hamish Gibbs, Joel Hellewell, James Munday, Patrick Barks, Paul Campbell, Flavio Finger, and Sebastian Funk. Covidregionaldata: Subnational data for the covid-19 outbreak. -, -(-): –, 2020. doi: 10.5281/zenodo.3957539.

[3] Samir Bhatt, Neil Ferguson, Seth Flaxman, Axel Gandy, Swapnil Mishra, and James A Scott. Semi-Mechanistic Bayesian modeling of COVID-19 with Renewal Processes. *Journal of the Royal Statistical Society Series A: Statistics in Society*, page qnad030, February 2023. ISSN 0964-1998. doi: 10.1093/jrsssa/qnad030.

[4] epiforecasts.io/covid. Covid-19: Temporal variation in transmission during the COVID-19 outbreak. https://epiforecasts.io/covid/, 2020.

[5] Tapiwa Ganyani, Cécile Kremer, Dongxuan Chen, Andrea Torneri, Christel Faes, Jacco Wallinga, and Niel Hens. Estimating the generation interval for coronavirus disease (COVID-19) based on symptom onset data, March 2020. *Euro Surveillance: Bulletin Europeen Sur Les Maladies Transmissibles = European Communicable Disease Bulletin*, 25(17):2000257, April 2020. ISSN 1560-7917. doi: 10.2807/1560-7917.ES.2020.25.17.2000257.

[6] Stephen A. Lauer, Kyra H. Grantz, Qifang Bi, Forrest K. Jones, Qulu Zheng, Hannah R. Meredith, Andrew S. Azman, Nicholas G. Reich, and Justin Lessler. The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application. *Annals of Internal Medicine*, pages M20–0504, March 2020. ISSN 0003-4819. doi: 10.7326/M20-0504.

[7] Gabriel Riutort-Mayol, Paul-Christian Bürkner, Michael R. Andersen, Arno Solin, and Aki Vehtari. Practical Hilbert space approximate Bayesian Gaussian processes for probabilistic programming. *Statistics and Computing*, 33(1): 17, December 2022. ISSN 1573-1375. doi: 10.1007/s11222-022-10167-2.

[8] James A. Scott, Axel Gandy, Swapnil Mishra, Juliette Unwin, Seth Flaxman, and Samir Bhatt. epidemia: Modeling of epidemics using hierarchical bayesian models, 2020. URL `https://imperialcollegelondon.github.io/epidemia/`. R package version 1.0.0.

[9] Katharine Sherratt, Sam Abbott, Sophie R. Meakin, Joel Hellewell, James D. Munday, Nikos Bosse, CMMID Covid-19 working Group, Mark Jit, and Sebastian Funk. Exploring surveillance data biases when estimating the reproduction number: With insights into subpopulation transmission of Covid-19 in England. page 2020.10.18.20214585, March 2021. doi: 10.1101/2020.10.18.20214585.

[10] Stan Development Team. RStan: the R interface to Stan, 2023. URL `https://mc-stan.org/`. R package version 2.21.8.

[11] Bo Xu, Bernardo Gutierrez, Sarah Hill, Samuel Scarpino, Alyssa Loskill, Jessie Wu, Kara Sewalk, Sumiko Mekaru, Alexander Zarebski, Oliver Pybus, David Pigott, and Moritz Kraemer. Epidemiological data from the nCoV-2019 outbreak: Early descriptions from publicly available data. http://virological.org/t/epidemiological-data-from-the-ncov-2019-outbreak-early-descriptions-from-publicly-available-data/337, 2020.

7 Discussion

7.1 Summary and contributions to existing work

The work presented in this thesis has made several contributions related to forecast evaluation and human judgement forecasting in epidemiology.

Forecast evaluation

The first major contribution is towards improving the evaluation of infectious disease forecasts, both from a practical and a theoretical perspective.

scoringutils

The `scoringutils` package helps make forecast evaluation and the necessary tools more accessible to researchers and decision makers. The package facilitates the evaluation of probabilistic forecasts using proper scoring rules by providing a general-purpose tool and a flexible framework. It is also the first package to offer extensive support for evaluating probabilistic forecasts in the form of predictive quantiles. This quantile-based format was used throughout this thesis and is a format used by several infectious disease Forecast Hubs (Reich et al., 2019; Cramer et al., 2022; Sherratt et al., 2022; Bracher et al., 2022). `scoringutils` improves usability over existing software for evaluating probabilistic forecasts. It is especially geared towards comparing forecasts from different models, regardless of how those forecasts were generated. All forecasts can be scored and summarised in a convenient `data.table` format and the package offers functionality to compare performance visually across different dimensions and account for missing forecasts. `scoringutils` has since been used in published and unpublished work supporting major public health organisations such as the US Centers for Disease Control and Prevention, the European Centre for Disease Prevention and Control, the UK Health Security Agency, and Médecins Sans Frontièrs.

Scoring epidemiological forecasts on transformed scales

In addition to improving the tools available for forecast evaluations, this thesis made a theoretical contribution to our understanding of forecast evaluations in the context of epidemiology. It argued that transforming forecasts and observations prior to applying the continuous ranked probability score (CRPS) or the weighted interval score (WIS) made it possible to obtain more meaningful results than with the CRPS and WIS values based on untransformed forecasts. Both the WIS and the CRPS measure the absolute distance between the forecast and the observation. Scores therefore tend to increase with the order of magnitude of the target and do not take the exponential nature of infectious disease

processes into account. This can be mitigated by transforming the forecasts and observations before scoring. The natural logarithm is a particularly attractive transformation in an epidemiological context, as the resulting score represents a measure of how well the forecasts predicted the exponential growth rate. The CRPS applied to log-transformed forecasts can also be understood as an approximate probabilistic version of the symmetric relative error. Furthermore, the natural log transformation can serve as a variance-stabilising transformation, helping to make scores more comparable across time, locations and forecast targets. We compared scores on the natural scale and on the log scale for a set of forecasts submitted to the European COVID-19 Forecast Hub and found that rankings between models changed. Scores were more evenly distributed across time, locations and forecast targets on the log scale. Forecasters were less severely penalised for missing the peak on the log scale and received higher penalties for missing an upswing of incidences.

Human judgement

The second major contribution of this thesis is in improving our understanding of the role of human judgement in infectious disease forecasting and the potential and limitations of human judgement forecasts.

In order to be able to compare probabilistic forecasts from human forecasters against computational models, we developed a new open source platform and R package, `crowdforecastr`, which allows the elicitation of probabilistic time series forecasts from individual forecasters. When we submitted our forecasts to the German and Polish Forecast Hub in 2020 and 2021, our study was the first one in an epidemiological context to compare full predictive distributions from individual users against model-based predictions. A previous study by Farrow et al. (2017) had elicited point forecasts from individual participants and combined them to a probabilistic forecast based on the mean and variation in participants' predictions.

In the first study presented in Chapter 4, we compared a small crowd of human forecasters against two minimally-tuned epidemiological models and an ensemble of model-based predictions submitted to the German and Polish Forecast Hub. These submissions contributed to a shared effort to inform the public and public health decision makers in Germany and Poland during the earlier phases of COVID-19. We found predictions for cases from human forecasters to be slightly better in terms of the weighted interval score than forecasts of the Hub ensemble, but worse when predicting deaths. Our minimally-tuned model forecasts performed comparable to our crowd forecasts for short horizons but noticeably worse for longer horizons. This suggests that human judgement is beneficial in guiding model-based predictions when conditions change over time. We also found that even forecasts that are worse than a pre-existing ensemble can help to improve ensemble forecasts when including it in that ensemble.

In the second study in the UK presented in Chapter 6, we repeated the basic set-up of the first study with a larger number of participants in a different country. Forecasts were elicited as part of a public forecasting challenge over 13 weeks and submitted to the European COVID-19 Forecast Hub. In addition, we tested a novel forecasting approach in which users submitted a forecast of the effective reproduction number R_t, which would then get mapped to reported cases and deaths using the R package `EpiNow2` (Abbott et al., 2020). Following

the conclusions presented in Chapter 5 we evaluated forecasts using both untransformed and log-transformed predictions and observations. We found the performance of human and model-based forecasts to be overall comparable. The Hub ensemble performed slightly better than human forecasts on the natural scale, and slightly worse when evaluated on the log scale. Our novel R_t approach performed comparably to other forecasts when predicting cases (which forecasters could observe directly), and noticeably worse for death predictions (which forecasters could not see), suggesting that the underlying model did not accurately capture the relation between cases and deaths.

7.2 Limitations

The work presented in this thesis is subject to various important limitations. This section gives an overview and summary of those limitations, while further details can be found in the corresponding chapters.

Forecast evaluation

The `scoringutils` package is still under active development and is not yet at the point where it is a fully general forecast evaluation package that can satisfy all needs a typical practitioner might have. One particular area of improvement is the number of forecast types that are supported by the package, which is currently limited. For example, the package currently does not allow users to evaluate forecasts in a binned format or predictions expressed in the form of closed-form distributions. It also does not support categorical forecasts and a variety of classification tasks that are common in many fields. The package also currently lacks a range of functionality, in particular with respect to statistical testing. It also lacks some visualisations, particularly related to model calibration that are available in other packages. While the package is aimed to be user-friendly, it still lacks some documentation as well as vignettes with additional explanations and case studies that make the package easily accessible to new users.

In terms of the theoretical advancements regarding transformations of forecasts before scoring, more work needs to be done, both with respect to transformations in general and the log-transformation in particular. Theoretical considerations suggest that log-transforming forecasts before applying the CRPS or WIS yields a score that better reflects the exponential nature of infectious disease processes. However, whether or not scores on the log scale appear indeed more meaningful to researchers and policy makers in practice remains to be seen. So far we have only conducted a small case study in order to illustrate the effects of log-transforming forecasts before evaluation. More work is needed to obtain a more complete understanding of the behaviour of scores on the log scale across different applications. Transformations in general may be a promising way of obtaining scores that are more meaningful to researchers and policy makers, but more work is needed to better understand when to use which transformation.

Several practical issues arise when transforming forecasts. The natural logarithm, for example, does not allow any negative or zero values in the forecasts or observations. In Chapter 5

we suggested to add a small value to all observations and forecasts in order to deal with zero values. This does not break propriety, but introduces more degrees of freedom and changes the scores in subtle ways. How to best deal with negative values remains unclear. Negative values may not be a great problem in an epidemiological context, where observations are usually counts, but could hinder application in other contexts. Issues may arise even when no zeros or negative values are present, especially when forecasts or observations are small. A small absolute difference between forecast and observation can translate to a large relative difference, causing scores to blow up. This issue can arise in particular if forecasts and observations are restricted to assume integer values. Users may find that forecasts made for small targets can dominate overall scores in an undesirable way, depending on the relationship between the mean and the variance of the forecast target. Bracher et al. (2021a) argued before that the fact that CRPS scales with the forecast target conveys meaningful information that gets lost when log-transforming targets. Some desirable transformations other than the natural logarithm, such as converting absolute forecasts to forecasts of weekly growth rates by dividing every predicted value by the value in the week before, are restricted to forecasts that are stored in a format that allows to trace the predictive distribution over time (such as storing sample forecast trajectories).

Human judgement

This thesis has studied the potential of human judgement to forecast infectious diseases such as COVID-19. Both studies, however, suffered from a low number of participants. It therefore remains somewhat unclear how well results could generalise to other settings.

The first study in Germany and Poland, presented in Chapter 4, had a median of six participants per week (the Hub ensemble had a similar amount of ensemble models). Participants were also recruited in a very ad hoc fashion among friends and colleagues, making the sample not representative. In this sense, the study is maybe better understood as a case study of an acute outbreak response effort using human judgement forecasts. The overall evaluation was made difficult by the fact that scores varied a lot from week to week and across locations and forecast horizons. This introduced many researcher degrees of freedom, as results and interpretations could change depending on how forecasts were evaluated. For the study in Germany and Poland, we have only looked at forecasts on the natural scale and therefore lack knowledge of what results would have looked like on the log scale.

Given the very context-dependent nature of human judgement forecasting, results may not generalise well to other settings. While we attempted to replicate some of our findings in Germany and Poland in a second study, there are a few factors that make a direct comparison difficult. In the first study, human forecasts were combined using a mean ensemble, while in the second study, we used a quantile-wise median (following the practice adopted by all COVID-19 Forecast Hubs). Changes in the number and composition of forecasters, the shorter time horizon and the different setting (including vaccination and a different COVID-19 variant) also limit comparability. In both studies, human forecasts were compared against an ensemble of model-based predictions. However, those ensembles differed both in terms of their size and their composition.

All forecasts were only analysed at an aggregate level. We therefore could not obtain an

understanding of how individuals contributed to the overall ensemble forecasts. For example, we cannot know how the number of participants influenced overall performance, or whether participants learned over time. Results may have looked different had we successfully retained participants throughout the studies. In both our studies, most participants only submitted a single forecast, which may have affected overall performance. In addition, we treated forecasters more or less as black boxes, without qualitatively investigating their thought processes in detail.

We asked forecasters to self-identify as "experts" by whether or not they worked in infectious disease modelling or had professional experience in any related field. However, the value of that information was limited in a few ways. Firstly, we were not able to check participants' statements. Secondly, the question we asked was perhaps too broad to provide a useful measure of the participant's actual expertise for the task at hand. Something like a short quiz may have helped to get more detailed information on the participants' level of expertise.

One aim of the study in Germany and Poland was to improve our understanding of the relative contributions of human judgement forecasting and epidemiological modelling. However, our approach of comparing direct human forecasts with minimally-tuned epidemiological models was likely flawed and may not have entirely achieved this goal. Firstly, there was a partial, but not complete overlap between the researchers who designed our minimally-tuned models and the participants of the forecasting study. Had these two groups either been disjunct or identical, a fairer comparison would have been possible. This, however, would have come at the expense of an even further reduced number of participants to the point that only one or two forecasters might have been left on a given date. Furthermore, human forecasters were allowed to use any model they liked as an input and we therefore cannot make statements about the extent to which human forecasts were guided by epidemiological modelling. The notion of "minimally-tuned" we used in our study is very vague, making it unclear how much our models were actually guided by human judgement rather than being just a mathematical representation of our abstract understanding of infectious disease dynamics. Furthermore, we only used two minimally-tuned models for this study. It is unclear whether our models are able to represent a general class of "minimally-tuned epidemiological models", or whether they are just two specific models with particular strengths and weaknesses that may not generalise to other models.

For the second crowd forecasting study in the UK, we experimented with a novel forecasting approach that asked human forecasters to predict the effective reproduction number R_t which then got mapped to reported cases and deaths. Forecasters were able to see a preview of the case forecast implied by their R_t forecast, but could not see the corresponding prediction for deaths. Participants therefore had to rely entirely on the underlying model and were not able to adjust parameters like the CFR or the delay between cases and deaths. The specific results we obtained therefore depend very strongly on the specific model used and therefore do not necessarily reflect the overall potential of the proposed approach.

7.3 Implications and avenues for future work

Forecast evaluation

Broad interest in the `scoringutils` package suggests that developing and maintaining tools may be an effective way to contribute to the field of infectious disease forecasting. Unfortunately, academia provides too few mechanisms to incentivise and reward the development and maintenance of tools in a collaborative and sustainable way. While creating a new tool allows the authors to get some recognition through the publication of their work, there are too few ways to reward researchers for contributing to existing projects or to put effort into continuously developing software after publication. This promotes an ecosystem with a large number of disconnected tools that solve parts of a larger problem, but do not interface well with each other. In some sense, the forecast evaluation ecosystem is not much different and it would be better to have a single forecast evaluation package on which efforts could be concentrated. For `scoringutils`, one important challenge will be to turn the package into a community project that is supported by a larger group of users and contributors from different institutions and backgrounds. This would make sure that the package is maintained and developed in a sustainable way in close collaboration with those who actually use it. In terms of actual development, a broad range of further improvements and features would be useful. One important area is the addition of tools to determine whether two forecasts perform significantly differently. Another is the expansion of forecast types supported by the package. Currently planned are support for categorical forecasts, as well as support for multivariate forecasts in which a predictive distribution is jointly defined over multiple targets (e.g. locations or time points). Another interesting idea could be scoring forecasts against distributions (as opposed to only scoring against observed values). This could be useful, for example, to evaluate forecasts of quantities like the effective reproduction number R_t, which are never directly observable, against the final best estimate available later on. Another planned improvement is the integration of the package with other forecast evaluation and modelling packages. This means on the one hand creating helper functions to convert from and to the formats used by different packages. On the other hand, it means creating vignettes and case studies that explain in detail how to use `scoringutils` in combination with other packages.

Despite the widespread use of proper scoring rules, forecast evaluation is far from a solved problem and much work remains to be done.

Firstly, we only have a rudimentary understanding of how the resulting scores translate into "usefulness" of the forecasts. For example, in some contexts, such as forecasting hospital bed occupancy, it might be much worse to underpredict actual numbers than overpredict them, in a way that is not reflected by ordinary scores. Or it might be that a forecast that is biased but correctly predicts a trend is more useful than one that shows the wrong trend but is closer to the actual values and therefore receives a better score. If we show decision makers different forecasts, which score will choose the model that different decision makers describe as the most useful one?

Transforming forecasts before scoring is one promising way of approaching this issue. Further

research into scores on the natural vs. on the log scale is needed. For example, it would be interesting to investigate whether scores on the log scale tend to be more consistent over time than those on the natural scale. Furthermore, developing and exploring new transformations may prove very useful. Forecast transformations may be particularly attractive in combination with the possibility of creating composite scores as a linear combination of different proper scoring rules. One could for example create a new score that is a weighted combination of scores on the natural and on the log scale. Transformations may be particularly useful when forecasts are represented in the form of predictive samples with sample trajectories. This would allow, for example, evaluating the shape of the forecast trend line by dividing forecasts for horizon h by the forecasts for horizon $h - 1$. Another promising approach may be to develop custom proper scoring rules for specific applications using Bayes Acts (Brehmer and Gneiting, 2020) and general loss functions.

Secondly, we do not have a good understanding of what adequate baselines for comparisons are. For example, predicting whether or not a stone will fall is much easier than predicting whether or not a stock will rise. This is not reflected through scores that only compare the forecast and the outcome. This makes comparisons of forecasts across different settings, locations, times and forecast targets very difficult. Without an appropriate baseline to compare a forecast against, the raw score alone is somewhat meaningless.

Thirdly, forecast evaluations suffer from a large number of researcher degrees of freedom. Results change a lot depending on choices such as what metric to use, which forecast horizon to focus on, whether forecasts are transformed before scoring etc. This introduces a subjective element into the evaluation and makes results prone to motivated reasoning in a way that is masked by the apparent objectivity of numerical evaluations. To mitigate this, researchers should aim to show a broad range of different scores and metrics. It could be a good idea to establish shared standards of what should be reported in a forecast evaluation and how. Other ideas include pre-registration of studies or interactive visualisations of scores that forecast consumers can explore on their own.

Fourthly, we still do not have a very good understanding of how much to trust a forecaster. In the context of the COVID-19 Forecast Hubs, models often showed consistently good performance for a long time and then suddenly broke down when circumstances changed. The Hub ensembles robustly emerged as good (and in most instances the best) choice on average across several Forecast Hubs and many evaluations. But even the ensemble often missed changes in trends and performed poorly. Many key questions still remain open: How much data is needed until we can say confidently that a model/forecaster performs well or badly? Can we identify some kind of reliability measure that would indicate how much we can trust a given forecast at a particular time? Should we use significance tests to compare models/forecasters and if so how much can we trust them? One issue with significance tests in particular is that forecasts are usually correlated across time and location, reducing the effective sample size. Also, as Diebold (2015) note, there is a difference between comparing forecasts with comparing forecasters, as the observed performance of a single set of forecasts is not necessarily representative of the performance of that forecaster in general.

Fifthly, the high dimensionality of forecasting data poses a problem for forecast evaluation. In an epidemiological setting, forecasts are often made by several models at different time points

for different forecast horizons, locations and forecast targets. Evaluating and visualising forecasts for different stratification of the data is cumbersome, difficult, and relies heavily on the researcher's judgement. It would be helpful to come up with robust ways to handle the dimensionality of the data better. One obvious candidate is modelling scores using a regression framework. A major obstacle, however, is that scores (at least in epidemiology) tend to be heavily skewed and dominated by outliers. Perhaps a combination of transforming forecasts before scoring and standardising scores may make it possible to use a regression framework, even if p-values for coefficients may not be reliable.

Finally, we do not have established good ways to provide useful feedback for forecasters on how to improve their forecasts. The Forecast Hubs published data on scores and average performance but struggled to provide actionable feedback. One option would be to create single-model evaluations that explore in detail when and how a model performed well or badly. A more technically sophisticated option might be to provide forecasters with an interface that allows them to manipulate past forecasts and observe how e.g. shifting forecasts or adjusting dispersion would have influenced past scores.

Human judgement forecasting

This thesis has attempted to shed some light on the opportunities and limitations associated with human judgement forecasting of infectious diseases such as COVID-19. Our studies provided some evidence that a mixed crowd of human forecasters and an ensemble of model-based predictions can produce forecasts of comparable quality. However, many of the details of what drives good performance, both in models and human forecasters, are still poorly understood.

Human judgement forecasting and mathematical modelling differ in their advantages and disadvantages. Whether or not one should aim to elicit forecasts from humans or mathematical models therefore depends mainly on for what purpose and in which circumstances forecasts are to be used. Eliciting both human judgement forecasts and model-based predictions for the same thing is useful to obtain an understanding of the relative performance and relative strength of the two different approaches. Going forward, however, obtaining forecasts for the exact same targets may not be the best use of resources. Rather, an important question in the future is how to elicit forecasts in a way that plays to the respective strengths of humans and mathematical models. Human judgement may be particularly useful in situations where we are interested in questions that cannot easily captured by modelling or where the relevant information is hard to feed into models. Human judgement forecasting requires a lot of effort though, and past studies often struggled to retain participants over a long time. Both our studies suffered from low participation. For the first study, the low number of participants was in part due to time constraints - the first forecasts were submitted within three weeks of the start of this PhD. However, even with more time and preparation, as was possible for the second study in the UK, we found it hard to recruit and retain a large number of participants. Generating forecasts over a prolonged period may be easier using mathematical models that can be automated and only have to be adapted occasionally when circumstances change. Mathematical modelling may also be advantageous in settings with good data quality and for forecasting tasks that benefit from the ability to do complex calculations. Another advantage

of mechanistic models is that they may help us to understand the underlying infectious disease process better and apply those learnings to other settings.

One interesting question that arose in this thesis was whether humans might have an advantage when predicting cases, whereas computational models might be better at forecasting a lagged quantity such as deaths. This was suggested by the results from our first study in Germany, as well as by McAndrew et al. (2022b). Our second study in the UK did not confirm this (but also did not provide strong evidence against this hypothesis). However, performance when forecasting deaths in our second study may have been affected by changes in the age composition of cases and the rise of the Delta variant in the UK. It could be possible that humans found it easier to adapt to changes in the case fatality ratio than models, possibly implying that humans generally tend to be good at incorporating uncertain information and adapting to novel circumstances. Computational models, on the other hand, could be better at estimating the delay distribution between cases and deaths precisely, if circumstances stay constant. This 'story' seems compelling, which is exactly why we should treat it with caution. Given the small number of studies and low sample sizes, we cannot draw strong conclusions and there is a high risk of overinterpreting noise. Ultimately, this highlights that we do not really have a sufficiently good understanding of how and why different forecasting approaches perform well.

One straightforward way to obtain a better understanding of human judgement forecasting would be to conduct qualitative interviews with forecasters. This could shed light on which factors forecasters take into account and the reasoning they apply. It could also help us understand in how far the reasoning of experts and non-experts differs and provide information on how to improve predictive performance in the future.

Qualitative interviews could be used not only to obtain insights on human judgement forecasting but also to get a better understanding of the role of human judgement in infectious disease modelling more generally. In our original study in Germany and Poland we described the output of mathematical modelling as a mixture between abstract mathematical assumptions and the subjective human judgement of the researchers developing the models. One aim of the study was to obtain a better understanding of the influence and effect of that human judgement has in the model development process. As discussed in the last section on limitations, it is questionable whether we achieved this to the extent we had aimed for. One way to address this question more directly would have been to conduct qualitative interviews with modellers who submitted models to the COVID-19 Forecast Hubs.

Another interesting way of obtaining a better understanding of the influence of human judgement in infectious disease forecasting would be to ask humans to directly manipulate model outputs. This could either be done with the researchers themselves who developed the models or a different set of forecasters. It would then be possible to analyse how human judgement alone, model-based predictions alone, or a combination of the two would perform. This would address our original research question more directly and could lead to promising ways of improving on the accuracy of both human and model-based predictions. Importantly, general-purpose tools that would allow users to adapt arbitrary forecasts are not readily available at present. This is somewhat astonishing, given the widespread use of forecasting and the fact that researchers, at least anecdotally, often meaningfully disagree with aspects

or details of their own model-based predictions.

Combining human judgement and model-based predictions in order to mitigate weaknesses of any one approach is appealing. Related efforts generally follow two aims: improving predictive performance, and making forecasting efforts more scalable by reducing the cognitive load for humans. Several promising ideas have been proposed. One, having human forecasters manually adapt model outputs, was mentioned in the last paragraph. A second one, asking a human forecaster to predict a quantity such as R_t or the growth rate of an infectious disease process, was discussed in our paper on the UK Crowd Forecasting Challenge in Chapter 6. While performance in our study was poor, users quite liked the interface. The study explored an early prototype and there are several improvements that could be made, such as allowing forecasters to see the implied forecasts for deaths and influence it through adjusting the CFR. Another approach is to use human judgement in order to estimate or forecast parameters that are then used as an input to mathematical models (see e.g. Venkatramanan et al., 2022). This, however, may not be easy as interpreting and estimating technical parameters such as the serial interval requires expert subject matter knowledge. One could, however, for example, ask participants to predict the timing and magnitude of a peak and use that information as a prior to constrain plausible model scenarios. Yet another idea would be to use human judgement to guide the formation of forecast ensembles. In the past, it has proven surprisingly difficult to create forecast ensembles that outperform a simple unweighted combination of individual forecasts (this has been called the "forecast combination puzzle", see e.g. Claeskens et al., 2016). Weighted ensembles are usually trained based on past performance. However, training algorithms usually lack a deeper understanding of the models involved. Organisers of the various COVID-19 Forecast Hubs, on the other hand, had quite a good understanding of the models involved and their respective strengths and weaknesses in various situations. Perhaps human judgement that is guided by an understanding of individual models could help in determining model weights that improve on equal weighting. Again, adequate tools would be needed to allow users to see the effect of choosing different model weights on the resulting ensemble. Somewhat related, it could be useful to ask forecasters (for example, researchers who submit forecasts to Forecast Hubs) for their personal confidence in their own forecasts. These subjective estimates might be used to improve the ensemble forecast or provide learning opportunities for participants.

7.4 Conclusion

This thesis has focused on the two areas of forecast evaluation and the comparison of human judgement forecasting and model-based predictions. Throughout, it has touched on a large variety of different questions and topics. What are appropriate tools to evaluate and elicit forecasts? How can we interpret different scores and make them meaningful and useful to modellers and policy makers alike? How can we create scores that better reflect the underlying infectious disease processes we are trying to predict? What are the strengths and weaknesses of human judgement forecasts and model-based predictions and how can we best cater to the respective strengths in the future?

This work has provided tangible progress on some of these questions. In particular with

regards to forecast evaluation, both `scoringutils` and the paper on forecast transformation (Chapter 5) represent important advances that have found their way into the work of major institutions. With regard to human judgement forecasting, this thesis has contributed to a shared understanding of the relative performance of humans and mathematical models in infectious disease forecasting and has explored several novel approaches to eliciting forecasts and combining human judgement and computational models.

Overall, likely more questions have been raised than answered. With regard to the human judgement forecasting parts of this thesis, the majority of the work should best be understood as exploratory. On the one hand, our work did contribute to existing knowledge both in terms of our understanding of the performance of human forecasts and the ways that forecast evaluations in general could be conducted. On the other hand, our two studies highlight how difficult it is to obtain conclusive answers, how many researcher degrees of freedom are involved and how many questions remain unsolved. In some sense, the major value of our work on comparing human and model-based forecasts perhaps comes from using a particular case study as a starting point for open exploration and reflection, and from raising interesting questions and suggesting novel areas of inquiry. With regard to the applied and theoretical work on forecast evaluation in this thesis, our efforts highlight how much there is still to be done in terms of interpreting the performance of forecasters. It remains difficult to link scores to the actual usefulness of a forecast. Furthermore, handling the high dimensionality of the different forecasts targets across time and location, and target types remains a challenge. On the other hand, the work on transforming forecasts before scoring them presented in this thesis opens up a large range of exciting possibilities and avenues for future research.

The work in this thesis, the tools and theoretical advancements, as well as the ideas and questions it raised hopefully contribute to an infectious disease forecasting ecosystem that allows us to be a little more prepared for the next infectious disease outbreaks than the last.

Bibliography

Abbott, S., Hellewell, J., Hickson, J., Munday, J., Gostic, K., Ellis, P., Sherratt, K., Gibbs, H., Thompson, R., Meakin, S., Bosse, N. I., Mee, P., and Funk, S. (2020). Epinow2: Estimate real-time case counts and time-varying epidemiological parameters. -, -(-):–.

Arvan, M., Fahimnia, B., Reisi, M., and Siemsen, E. (2019). Integrating human judgement into quantitative forecasting methods: A review. *Omega*, 86:237–252.

Atanasov, P., Rescober, P., Stone, E., Swift, S. A., Servan-Schreiber, E., Tetlock, P., Ungar, L., and Mellers, B. (2016). Distilling the Wisdom of Crowds: Prediction Markets vs. Prediction Polls. *Management Science*, 63(3):691–706.

Bernice Brown (1968). DELPHI PROCESS: A METHODOLOGY USED FOR THE ELICITATION OF OPINIONS OF EXPERTS. Technical report.

Bhatt, S., Ferguson, N., Flaxman, S., Gandy, A., Mishra, S., and Scott, J. A. (2023). Semi-mechanistic Bayesian modelling of COVID-19 with renewal processes. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 186(4):601–615.

Biggerstaff, M., Alper, D., Dredze, M., Fox, S., Fung, I. C.-H., Hickmann, K. S., Lewis, B., Rosenfeld, R., Shaman, J., Tsou, M.-H., Velardi, P., Vespignani, A., Finelli, L., and for the Influenza Forecasting Contest Working Group (2016). Results from the centers for disease control and prevention's predict the 2013–2014 Influenza Season Challenge. *BMC Infectious Diseases*, 16(1):357.

Bosse, N. I., Abbott, S., Bracher, J., Hain, H., Quilty, B. J., Jit, M., Group, C. f. t. M. M. o. I. D. C.-. W., van Leeuwen, E., Cori, A., and Funk, S. (2022a). Comparing human and model-based forecasts of COVID-19 in Germany and Poland. *PLOS Computational Biology*, 18(9):e1010405.

Bosse, N. I., Abbott, S., Cori, A., van Leeuwen, E., Bracher, J., and Funk, S. (2023a). Transformation of forecasts for evaluating predictive performance in an epidemiological context.

Bosse, N. I., Abbott, S., EpiForecasts, and Funk, S. (2020). *Crowdforecastr: eliciting crowd forecasts in R shiny*.

Bosse, N. I., Gruson, H., Cori, A., van Leeuwen, E., Funk, S., and Abbott, S. (2022b). Evaluating Forecasts with scoringutils in R. *arXiv*.

Bosse, NI., Abbott, S., Bracher, J., van Leeuwen, E., Cori, A., and Funk, S. (2023b). Human judgement forecasting of COVID-19 in the UK [version 1; peer review: 1 approved with reservations]. *Wellcome Open Research*, 8(416).

Box, G. E. P. and Jenkins, G. M. (1970). *Time Series Analysis: Forecasting and Control*. Holden-Day.

Bracher, J., Ray, E. L., Gneiting, T., and Reich, N. G. (2021a). Evaluating epidemic forecasts in an interval format. *PLoS computational biology*, 17(2):e1008618.

Bracher, J., Wolffram, D., Deuschel, J., Görgen, K., Ketterer, J. L., Ullrich, A., Abbott, S., Barbarossa, M. V., Bertsimas, D., Bhatia, S., Bodych, M., Bosse, N. I., Burgard, J. P., Castro, L., Fairchild, G., Fiedler, J., Fuhrmann, J., Funk, S., Gambin, A., Gogolewski, K., Heyder, S., Hotz, T., Kheifetz, Y., Kirsten, H., Krueger, T., Krymova, E., Leithäuser, N., Li, M. L., Meinke, J. H., Miasojedow, B., Michaud, I. J., Mohring, J., Nouvellet, P., Nowosielski, J. M., Ozanski, T., Radwan, M., Rakowski, F., Scholz, M., Soni, S., Srivastava, A., Gneiting, T., and Schienle, M. (2022). National and subnational short-term forecasting of COVID-19 in Germany and Poland during early 2021. *Communications Medicine*, 2(1):1–17.

Bracher, J., Wolffram, D., Deuschel, J., Görgen, K., Ketterer, J. L., Ullrich, A., Abbott, S., Barbarossa, M. V., Bertsimas, D., Bhatia, S., Bodych, M., Bosse, N. I., Burgard, J. P., Castro, L., Fairchild, G., Fuhrmann, J., Funk, S., Gogolewski, K., Gu, Q., Heyder, S., Hotz, T., Kheifetz, Y., Kirsten, H., Krueger, T., Krymova, E., Li, M. L., Meinke, J. H., Michaud, I. J., Niedzielewski, K., Ożański, T., Rakowski, F., Scholz, M., Soni, S., Srivastava, A., Zieliński, J., Zou, D., Gneiting, T., and Schienle, M. (2021b). A pre-registered short-term forecasting study of COVID-19 in Germany and Poland during the second wave. *Nature Communications*, 12(1):5173.

Brehmer, J. R. and Gneiting, T. (2020). Properization: Constructing proper scoring rules via Bayes acts. *Annals of the Institute of Statistical Mathematics*, 72(3):659–673.

Brier, G. W. (1950). Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review*, 78(1):1–3.

Brooks, L. C., Farrow, D. C., Hyun, S., Tibshirani, R. J., and Rosenfeld, R. (2018). Non-mechanistic forecasts of seasonal influenza with iterative one-week-ahead distributions. *PLOS Computational Biology*, 14(6):e1006134.

Claeskens, G., Magnus, J. R., Vasnev, A. L., and Wang, W. (2016). The forecast combination puzzle: A simple theoretical explanation. *International Journal of Forecasting*, 32(3):754–762.

Cramer, E. Y., Ray, E. L., Lopez, V. K., Bracher, J., Brennen, A., Castro Rivadeneira, A. J., Gerding, A., Gneiting, T., House, K. H., Huang, Y., Jayawardena, D., Kanji, A. H., Khandelwal, A., Le, K., Mühlemann, A., Niemi, J., Shah, A., Stark, A., Wang, Y., Wattanachit, N., Zorn, M. W., Gu, Y., Jain, S., Bannur, N., Deva, A., Kulkarni, M., Merugu, S., Raval, A., Shingi, S., Tiwari, A., White, J., Abernethy, N. F., Woody, S., Dahan, M., Fox, S., Gaither, K., Lachmann, M., Meyers, L. A., Scott, J. G., Tec, M., Srivastava, A., George, G. E., Cegan, J. C., Dettwiller, I. D., England, W. P., Farthing, M. W., Hunter, R. H., Lafferty, B., Linkov, I., Mayo, M. L., Parno, M. D., Rowland, M. A., Trump, B. D., Zhang-James, Y., Chen, S., Faraone, S. V., Hess, J., Morley, C. P., Salekin, A., Wang, D., Corsetti, S. M., Baer, T. M., Eisenberg, M. C., Falb, K., Huang, Y., Martin, E. T., McCauley, E., Myers, R. L., Schwarz, T., Sheldon, D., Gibson, G. C., Yu, R., Gao, L., Ma, Y., Wu, D., Yan, X., Jin, X., Wang, Y.-X., Chen, Y., Guo, L.,

Zhao, Y., Gu, Q., Chen, J., Wang, L., Xu, P., Zhang, W., Zou, D., Biegel, H., Lega, J., McConnell, S., Nagraj, V. P., Guertin, S. L., Hulme-Lowe, C., Turner, S. D., Shi, Y., Ban, X., Walraven, R., Hong, Q.-J., Kong, S., van de Walle, A., Turtle, J. A., Ben-Nun, M., Riley, S., Riley, P., Koyluoglu, U., DesRoches, D., Forli, P., Hamory, B., Kyriakides, C., Leis, H., Milliken, J., Moloney, M., Morgan, J., Nirgudkar, N., Ozcan, G., Piwonka, N., Ravi, M., Schrader, C., Shakhnovich, E., Siegel, D., Spatz, R., Stiefeling, C., Wilkinson, B., Wong, A., Cavany, S., España, G., Moore, S., Oidtman, R., Perkins, A., Kraus, D., Kraus, A., Gao, Z., Bian, J., Cao, W., Lavista Ferres, J., Li, C., Liu, T.-Y., Xie, X., Zhang, S., Zheng, S., Vespignani, A., Chinazzi, M., Davis, J. T., Mu, K., Pastore y Piontti, A., Xiong, X., Zheng, A., Baek, J., Farias, V., Georgescu, A., Levi, R., Sinha, D., Wilde, J., Perakis, G., Bennouna, M. A., Nze-Ndong, D., Singhvi, D., Spantidakis, I., Thayaparan, L., Tsiourvas, A., Sarker, A., Jadbabaie, A., Shah, D., Della Penna, N., Celi, L. A., Sundar, S., Wolfinger, R., Osthus, D., Castro, L., Fairchild, G., Michaud, I., Karlen, D., Kinsey, M., Mullany, L. C., Rainwater-Lovett, K., Shin, L., Tallaksen, K., Wilson, S., Lee, E. C., Dent, J., Grantz, K. H., Hill, A. L., Kaminsky, J., Kaminsky, K., Keegan, L. T., Lauer, S. A., Lemaitre, J. C., Lessler, J., Meredith, H. R., Perez-Saez, J., Shah, S., Smith, C. P., Truelove, S. A., Wills, J., Marshall, M., Gardner, L., Nixon, K., Burant, J. C., Wang, L., Gao, L., Gu, Z., Kim, M., Li, X., Wang, G., Wang, Y., Yu, S., Reiner, R. C., Barber, R., Gakidou, E., Hay, S. I., Lim, S., Murray, C., Pigott, D., Gurung, H. L., Baccam, P., Stage, S. A., Suchoski, B. T., Prakash, B. A., Adhikari, B., Cui, J., Rodríguez, A., Tabassum, A., Xie, J., Keskinocak, P., Asplund, J., Baxter, A., Oruc, B. E., Serban, N., Arik, S. O., Dusenberry, M., Epshteyn, A., Kanal, E., Le, L. T., Li, C.-L., Pfister, T., Sava, D., Sinha, R., Tsai, T., Yoder, N., Yoon, J., Zhang, L., Abbott, S., Bosse, N. I., Funk, S., Hellewell, J., Meakin, S. R., Sherratt, K., Zhou, M., Kalantari, R., Yamana, T. K., Pei, S., Shaman, J., Li, M. L., Bertsimas, D., Skali Lami, O., Soni, S., Tazi Bouardi, H., Ayer, T., Adee, M., Chhatwal, J., Dalgic, O. O., Ladd, M. A., Linas, B. P., Mueller, P., Xiao, J., Wang, Y., Wang, Q., Xie, S., Zeng, D., Green, A., Bien, J., Brooks, L., Hu, A. J., Jahja, M., McDonald, D., Narasimhan, B., Politsch, C., Rajanala, S., Rumack, A., Simon, N., Tibshirani, R. J., Tibshirani, R., Ventura, V., Wasserman, L., O'Dea, E. B., Drake, J. M., Pagano, R., Tran, Q. T., Ho, L. S. T., Huynh, H., Walker, J. W., Slayton, R. B., Johansson, M. A., Biggerstaff, M., and Reich, N. G. (2022). Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States. *Proceedings of the National Academy of Sciences*, 119(15):e2113561119.

Dalkey, N. and Helmer, O. (1963). An Experimental Application of the DELPHI Method to the Use of Experts. *Management Science*, 9(3):458–467.

Davies, N. and Ferris, S. (2022). Human judgement forecasting tournaments: A feasibility study based on the COVID-19 pandemic with public health practitioners in England. *Public Health in Practice*, 3:100260.

Diebold, F. X. (2015). Comparing Predictive Accuracy, Twenty Years Later: A Personal Perspective on the Use and Abuse of Diebold–Mariano Tests. *Journal of Business & Economic Statistics*, 33(1):1–1.

Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y., Nosek, B. A.,

and Johannesson, M. (2015). Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences*, 112(50):15343–15347.

Epstein, E. S. (1969). A Scoring System for Probability Forecasts of Ranked Categories. *Journal of Applied Meteorology*, 8(6):985–987.

Farrow, D. C., Brooks, L. C., Hyun, S., Tibshirani, R. J., Burke, D. S., and Rosenfeld, R. (2017). A human judgment approach to epidemiological forecasting. *PLOS Computational Biology*, 13(3):e1005248.

Frauenthal, J. C. (1980). *Mathematical Modeling in Epidemiology*. Universitext. Springer, Berlin, Heidelberg.

Funk, S., Abbott, S., Atkins, B. D., Baguelin, M., Baillie, J. K., Birrell, P., Blake, J., Bosse, N. I., Burton, J., Carruthers, J., Davies, N. G., Angelis, D. D., Dyson, L., Edmunds, W. J., Eggo, R. M., Ferguson, N. M., Gaythorpe, K., Gorsich, E., Guyver-Fletcher, G., Hellewell, J., Hill, E. M., Holmes, A., House, T. A., Jewell, C., Jit, M., Jombart, T., Joshi, I., Keeling, M. J., Kendall, E., Knock, E. S., Kucharski, A. J., Lythgoe, K. A., Meakin, S. R., Munday, J. D., Openshaw, P. J. M., Overton, C. E., Pagani, F., Pearson, J., Perez-Guzman, P. N., Pellis, L., Scarabel, F., Semple, M. G., Sherratt, K., Tang, M., Tildesley, M. J., van Leeuwen, E., Whittles, L. K., Group, C. C.-. W., Team, I. C. C.-. R., and Investigators, I. (2020). Short-term forecasts to inform the response to the Covid-19 epidemic in the UK. *medRxiv*, page 2020.11.11.20220962.

Gerding, A., Reich, N. G., Rogers, B., and Ray, E. L. (2024). Evaluating infectious disease forecasts with allocation scoring rules.

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268.

Gneiting, T. and Raftery, A. E. (2005). Weather Forecasting with Ensemble Methods. *Science*, 310(5746):248–249.

Gneiting, T. and Raftery, A. E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Good, I. J. (1952). Rational Decisions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 14(1):107–114.

Gordon, M., Viganola, D., Bishop, M., Chen, Y., Dreber, A., Goldfedder, B., Holzmeister, F., Johannesson, M., Liu, Y., Twardy, C., Wang, J., and Pfeiffer, T. (2020). Are replication rates the same across academic fields? Community forecasts from the DARPA SCORE programme. *Royal Society Open Science*, 7(7):200566.

Held, L., Meyer, S., and Bracher, J. (2017). Probabilistic forecasting in infectious disease epidemiology: The 13th Armitage lecture. *Statistics in Medicine*, 36(22):3443–3460.

Hunter, E., Mac Namee, B., and Kelleher, J. D. (2017). A Taxonomy for Agent-Based Models in Human Infectious Disease Epidemiology. *Journal of Artificial Societies and Social Simulation*, 20(3):2.

Hyndman, Rob J and Athanasopoulos, George (2021). *Forecasting: Principles and Practice*.

Johansson, M. A., Apfeldorf, K. M., Dobson, S., Devita, J., Buczak, A. L., Baugher, B., Moniz, L. J., Bagley, T., Babin, S. M., Guven, E., Yamana, T. K., Shaman, J., Moschou, T., Lothian, N., Lane, A., Osborne, G., Jiang, G., Brooks, L. C., Farrow, D. C., Hyun, S., Tibshirani, R. J., Rosenfeld, R., Lessler, J., Reich, N. G., Cummings, D. A. T., Lauer, S. A., Moore, S. M., Clapham, H. E., Lowe, R., Bailey, T. C., García-Díez, M., Carvalho, M. S., Rodó, X., Sardar, T., Paul, R., Ray, E. L., Sakrejda, K., Brown, A. C., Meng, X., Osoba, O., Vardavas, R., Manheim, D., Moore, M., Rao, D. M., Porco, T. C., Ackley, S., Liu, F., Worden, L., Convertino, M., Liu, Y., Reddy, A., Ortiz, E., Rivero, J., Brito, H., Juarrero, A., Johnson, L. R., Gramacy, R. B., Cohen, J. M., Mordecai, E. A., Murdock, C. C., Rohr, J. R., Ryan, S. J., Stewart-Ibarra, A. M., Weikel, D. P., Jutla, A., Khan, R., Poultney, M., Colwell, R. R., Rivera-García, B., Barker, C. M., Bell, J. E., Biggerstaff, M., Swerdlow, D., Mier-y-Teran-Romero, L., Forshey, B. M., Trtanj, J., Asher, J., Clay, M., Margolis, H. S., Hebbeler, A. M., George, D., and Chretien, J.-P. (2019). An open challenge to advance probabilistic forecasting for dengue epidemics. *Proceedings of the National Academy of Sciences*, 116(48):24268–24274.

Jordan, A., Krüger, F., and Lerch, S. (2019). Evaluating Probabilistic Forecasts with **scoringRules**. *Journal of Statistical Software*, 90(12).

Kretzschmar, M. and Wallinga, J. (2009). Mathematical Models in Infectious Disease Epidemiology. *Modern Infectious Disease Epidemiology*, pages 209–221.

Machete, R. L. (2012). Contrasting Probabilistic Scoring Rules. *arXiv:1112.4530 [math, stat]*.

Matheson, J. E. and Winkler, R. L. (1976). Scoring Rules for Continuous Probability Distributions. *Management Science*, 22(10):1087–1096.

McAndrew, T., Cambeiro, J., and Besiroglu, T. (2022a). Aggregating human judgment probabilistic predictions of the safety, efficacy, and timing of a COVID-19 vaccine. *Vaccine*, 40(15):2331–2341.

McAndrew, T., Codi, A., Cambeiro, J., Besiroglu, T., Braun, D., Chen, E., De Cèsaris, L. E. U., and Luk, D. (2022b). Chimeric forecasting: Combining probabilistic predictions from computational models and human judgment. *BMC infectious diseases*, 22(1):833.

McAndrew, T., Majumder, M. S., Lover, A. A., Venkatramanan, S., Bocchini, P., Besiroglu, T., Codi, A., Braun, D., Dempsey, G., Abbott, S., Chevalier, S., Bosse, N. I., and Cambeiro, J. (2022c). Early human judgment forecasts of human monkeypox, May 2022. *The Lancet Digital Health*, 4(8):e569–e571.

McAndrew, T. and Reich, N. G. (2022). An expert judgment model to predict early stages of the COVID-19 pandemic in the United States. *PLOS Computational Biology*, 18(9):e1010485.

McGowan, C. J., Biggerstaff, M., Johansson, M., Apfeldorf, K. M., Ben-Nun, M., Brooks, L., Convertino, M., Erraguntla, M., Farrow, D. C., Freeze, J., Ghosh, S., Hyun, S., Kandula, S., Lega, J., Liu, Y., Michaud, N., Morita, H., Niemi, J., Ramakrishnan, N., Ray, E. L.,

Reich, N. G., Riley, P., Shaman, J., Tibshirani, R., Vespignani, A., Zhang, Q., and Reed, C. (2019). Collaborative efforts to forecast seasonal influenza in the United States, 2015–2016. *Scientific Reports*, 9(1):683.

Murphy, A. H. (1971). A Note on the Ranked Probability Score. *Journal of Applied Meteorology*, 10(1):155–156.

Page, A., Potter, K., Clifford, R., McLachlan, A., and Etherton-Beer, C. (2015). Prescribing for Australians living with dementia: Study protocol using the Delphi technique. *BMJ Open*, 5(8):e008048.

Ramos, M. H., van Andel, S. J., and Pappenberger, F. (2013). Do probabilistic forecasts lead to better decisions? *Hydrology and Earth System Sciences*, 17(6):2219–2232.

Recchia, G., Freeman, A. L. J., and Spiegelhalter, D. (2021). How well did experts and laypeople forecast the size of the COVID-19 pandemic? *PLOS ONE*, 16(5):e0250935.

Reich, N. G., Brooks, L. C., Fox, S. J., Kandula, S., McGowan, C. J., Moore, E., Osthus, D., Ray, E. L., Tushar, A., Yamana, T. K., Biggerstaff, M., Johansson, M. A., Rosenfeld, R., and Shaman, J. (2019). A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States. *Proceedings of the National Academy of Sciences*, 116(8):3146–3154.

Shah, N. H. and Mittal, M. (2021). Introduction to Compartmental Models in Epidemiology. In Shah, N. H. and Mittal, M., editors, *Mathematical Analysis for Transmission of COVID-19*, pages 1–20, Singapore. Springer.

Sherratt, K., Gruson, H., Grah, R., Johnson, H., Niehus, R., Prasse, B., Sandman, F., Deuschel, J., Wolffram, D., Abbott, S., Ullrich, A., Gibson, G., Ray, EL., Reich, NG., Sheldon, D., Wang, Y., Wattanachit, N., Wang, L., Trnka, J., Obozinski, G., Sun, T., Thanou, D., Pottier, L., Krymova, E., Barbarossa, MV., Leithäuser, N., Mohring, J., Schneider, J., Wlazlo, J., Fuhrmann, J., Lange, B., Rodiah, I., Baccam, P., Gurung, H., Stage, S., Suchoski, B., Budzinski, J., Walraven, R., Villanueva, I., Tucek, V., Šmíd, M., Zajícek, M., Pérez, Á. C., Reina, B., Bosse, NI., Meakin, S., Di Loro, A., Maruotti, A., Eclerová, V., Kraus, A., Kraus, D., Pribylova, L., Dimitris, B., Li, ML., Saksham, S., Dehning, J., Mohr, S., Priesemann, V., Redlarski, G., Bejar, B., Ardenghi, G., Parolini, N., Ziarelli, G., Bock, W., Heyder, S., Hotz, T., E., S. D., Guzman-Merino, M., Aznarte, JL., Moriña, D., Alonso, S., Álvarez, E., López, D., Prats, C., Burgard, JP., Rodloff, A., Zimmermann, T., Kuhlmann, A., Zibert, J., Pennoni, F., Divino, F., Català, M., Lovison, G., Giudici, P., Tarantino, B., Bartolucci, F., Jona, L. G., Mingione, M., Farcomeni, A., Srivastava, A., Montero-Manso, P., Adiga, A., Hurt, B., Lewis, B., Marathe, M., Porebski, P., Venkatramanan, S., Bartczuk, R., Dreger, F., Gambin, A., Gogolewski, K., Gruziel-Slomka, M., Krupa, B., Moszynski, A., Niedzielewski, K., Nowosielski, J., Radwan, M., Rakowski, F., Semeniuk, M., Szczurek, E., Zielinski, J., Kisielewski, J., Pabjan, B., Holger, K., Kheifetz, Y., Scholz, M., Bodych, M., Filinski, M., Idzikowski, R., Krueger, T., Ozanski, T., Bracher, J., and Funk, S. (2022). Predictive performance of multi-model ensemble forecasts of COVID-19 across European nation. *Europe PMC*.

Tetlock, P. E., Mellers, B. A., Rohrbaugh, N., and Chen, E. (2014). Forecasting Tournaments: Tools for Increasing Transparency and Improving the Quality of Debate. *Current Directions in Psychological Science*, 23(4):290–295.

Venkatramanan, S., Cambeiro, J., Liptay, T., Lewis, B., Orr, M., Dempsey, G., Telionis, A., Crow, J., Barrett, C., and Marathe, M. (2022). Utility of human judgment ensembles during times of pandemic uncertainty: A case study during the COVID-19 Omicron BA.1 wave in the USA.

Wallinga, J. and Lipsitch, M. (2006). How generation intervals shape the relationship between growth rates and reproductive numbers. *Proceedings of the Royal Society B: Biological Sciences*, 274(1609):599–604.

West, M. and Harrison, J. (1997). *Bayesian Forecasting and Dynamic Models*. Springer Series in Statistics. Springer-Verlag, New York.

Winkler, R. L. (1972). A Decision-Theoretic Approach to Interval Estimation. *Journal of the American Statistical Association*, 67(337):187–191.

Winkler, R. L., Muñoz, J., Cervera, J. L., Bernardo, J. M., Blattenberger, G., Kadane, J. B., Lindley, D. V., Murphy, A. H., Oliver, R. M., and Ríos-Insua, D. (1996). Scoring rules and the evaluation of probabilities. *Test*, 5(1):1–60.

Yamana, T. K., Kandula, S., and Shaman, J. (2016). Superensemble forecasts of dengue outbreaks. *Journal of The Royal Society Interface*, 13(123):20160410.

Ziel, F. (2021). The energy distance for ensemble and scenario reduction. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379(2202):20190431.