Predictive performance of multi-model ensemble forecasts of COVID-19 across European nations

Sherratt, K. [1], Gruson, H. [1], Grah, R. [2], Johnson, H. [2], Niehus, R. [2], Prasse, B. [2], Sandmann, F. [2], Deuschel, J. [3], Wolffram, D. [3], Abbott, S. [1], Ullrich, A. [4], Gibson, G. [5], Ray, EL. [5], Reich, NG. [5], Sheldon, D. [5], Wang, Y. [5], Wattanachit, N. [5], Wang, L. [6], Trnka, J. [7], Obozinski, G. [8], Sun, T. [8], Thanou, D. [8], Pottier, L. [9], Krymova, E. [10], Meinke, JH. [11], Barbarossa, MV. [12], Leithäuser, N. [13], Mohring, J. [13], Schneider, J. [13], Właźlo, J. [13], Fuhrmann, J. [14], Lange, B. [15], Rodiah, I. [15], Baccam, P. [16], Gurung, H. [16], Stage, S. [16], Suchoski, B. [16], Budzinski, J. [17], Walraven, R. [17], Villanueva, I. [18], Tuček, V. [19], Šmíd, M. [20], Zajíček, M. [20], Pérez Álvarez, C. [21], Reina, B. [21], Bosse, NI. [1], Meakin, S. [1], Castro, L. [22], Fairchild, G. [22], Michaud, I. [22], Osthus, D. [22], Alaimo Di Loro, P. [23], Maruotti, A. [23], Eclerová, V. [24], Kraus, A. [24], Kraus, D. [24], Pribylova, L. [24], Dimitris, B. [25], Li, ML. [25], Saksham, S. [25], Dehning, J. [26], Mohr, S. [26], Priesemann, V. [26], Redlarski, G. [27], Bejar, B. [28], Ardenghi, G. [29], Parolini, N. [29], Ziarelli, G. [29], Bock, W. [30], Heyder, S. [31], Hotz, T. [31], E. Singh, D. [32], Guzman-Merino, M. [32], Aznarte, JL. [33], Moriña, D. [34], Alonso, S. [35], Álvarez, E. [35], López, D. [35], Prats, C. [35], Burgard, JP. [36], Rodloff, A. [37], Zimmermann, T. [37], Kuhlmann, A. [38], Zibert, J. [39], Pennoni, F. [40], Divino, F. [41], Català, M. [42], Lovison, G. [43], Giudici, P. [44], Tarantino, B. [44], Bartolucci, F. [45], Jona Lasinio, G. [46], Mingione, M. [46], Farcomeni, A. [47], Srivastava, A. [48], Montero-Manso, P. [49], Adiga, A. [50], Hurt, B. [50], Lewis, B. [50], Marathe, M. [50], Porebski, P. [50], Venkatramanan, S. [50], Bartczuk, R. [51], Dreger, F. [51], Gambin, A. [51], Gogolewski, K. [51], Gruziel-Słomka, M. [51], Krupa, B. [51], Moszynski, A. [51], Niedzielewski, K. [51], Nowosielski, J. [51], Radwan, M. [51], Rakowski, F. [51], Semeniuk, M. [51], Szczurek, E. [51], Zieliński, J. [51], Kisielewski, J. [52], Pabjan, B. [53], Holger, K. [54], Kheifetz, Y. [54], Scholz, M. [54], Biecek, P. [55], Bodych, M. [56], Filinski, M. [56], Idzikowski, R. [56], Krueger, T. [56], Ozanski, T. [56], Bracher, J. [3], Funk, S. [1]

[1]London School of Hygiene & Tropical Medicine, [2]European Centre for Disease Prevention and Control (ECDC), [3]Karlsruhe Institute of Technology, [4]Robert Koch Institute, [5]University of Massachusetts Amherst, [6]Boston Children's Hospital and Harvard Medical School, [7]Charles University, [8]École Polytechnique Fédérale de Lausanne, [9]Éducation nationale, [10]Eidgenössische Technische Hochschule Zürich, [11]Forschungszentrum Jülich GmbH, [12]Frankfurt Institute for Advanced Studies, [13]Fraunhofer Institute for Industrial Mathematics,

[14]Heidelberg University, [15]Helmholtz Centre for Infection Research, [16]IEM, Inc., [17]Independent, [18]Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Universitat Pompeu Fabra, [19]Institute of Computer Science of the CAS, [20]Institute of Information Theory and Automation of the CAS, [21]Inverence, [22]Los Alamos National Laboratory, [23]LUMSA University, [24]Masaryk University, [25]Massachusetts Institute of Technology, [26]Max-Planck-Institut für Dynamik und Selbstorganisation, [27]Medical University of Gdansk, [28]Paul Scherrer Institute, [29]Politecnico di Milano, [30]Technical University of Kaiserlautern, [31]Technische Universität Ilmenau, [32]Universidad Carlos III de Madrid, [33]Universidad Nacional de Educación a Distancia (UNED), [34]Universitat de Barcelona, [35]Universitat Politècnica de Catalunya, [36]Universität Trier, [37]University of Cologne, [38]University of Halle, [39]University of Ljubljana, [40]University of Milano-Bicocca, [41]University of Molise, [42]University of Oxford, [43]University of Palermo, [44]University of Pavia, [45]University of Perugia, [46]University of Rome "La Sapienza", [47]University of Rome "Tor Vergata", [48]University of Southern California, [49]University of Sydney, [50]University of Virginia, [51]University of Warsaw, [52]University of Warsaw, University of Bialystok, [53]University of Wroclaw, [54]Universtät Leipzig, [55]Warsaw University of Technology, [56]Wroclaw University of Science and Technology

Complete author funding information available as a supplement

Abstract

Background: Short-term forecasts of infectious disease burden can contribute to situational awareness and aid capacity planning. Based on best practice in other fields and recent insights in infectious disease epidemiology, one can maximise the predictive performance of such forecasts if multiple models are combined into an ensemble. Here we report on the performance of ensembles in predicting COVID-19 cases and deaths across Europe between 08 March 2021 and 07 March 2022.

Methods: We used open-source tools to develop a public European COVID-19 Forecast Hub. We invited groups globally to contribute weekly forecasts for COVID-19 cases and deaths reported by a standardised source for 32 countries over the next one to four weeks. Teams submitted forecasts from March 2021 using standardised quantiles of the predictive distribution. Each week we created an ensemble forecast, where each predictive quantile was calculated as the equally-weighted average (initially the mean and then from 26th July the median) of all individual models' predictive quantiles. We measured the performance of each model using the relative Weighted Interval Score (WIS), comparing models' forecast accuracy relative to all other models. We retrospectively explored alternative methods for ensemble forecasts, including weighted averages based on models' past predictive performance.

Results: Over 52 weeks we collected forecasts from 48 unique models. We evaluated 29 models' forecast scores in comparison to the ensemble model. We found a weekly ensemble had a consistently strong performance across countries over time. Across all horizons and locations, the ensemble performed better on relative WIS than 83% of participating models' forecasts of incident cases (with a total N=886 predictions from 23 unique models), and 91% of participating models' forecasts of deaths (N=763 predictions from 20 models). Across a one to four week time horizon, ensemble performance declined with longer forecast periods when forecasting cases, but remained stable over four weeks for incident death forecasts. In every forecast across 32 countries, the ensemble outperformed most contributing models when forecasting either cases or deaths, frequently outperforming all of its individual component models. Among several choices of ensemble methods we found that the most influential and best choice was to use a median average of models instead of using the mean, regardless of methods of weighting component forecast models.

Conclusions: Our results support the use of combining forecasts from individual models into an ensemble in order to improve predictive performance across epidemiological targets and populations during infectious disease epidemics. Our findings further suggest that median ensemble methods yield better predictive performance more than ones based on means. Our findings also highlight that forecast consumers should place more weight on incident death forecasts than incident case forecasts at forecast horizons greater than two weeks.

Code and data availability: All data and code are publicly available on Github: covid19-forecast-hub-europe/euro-hub-ensemble.

Background

Epidemiological forecasts make quantitative statements about a disease outcome in the near future. Forecasting targets can include measures of prevalent or incident disease and its severity, for some population over a specified time horizon. Researchers, policy makers, and the general public have used such forecasts to understand and respond to the global outbreaks of COVID-19 [1]–[3]. At the same time, forecasters use a variety of methods and models for creating and publishing forecasts, varying in both defining the forecast outcome and in reporting the probability distribution of outcomes [4], [5].

Within Europe, comparing forecasts across both models and countries can support a range of national policy needs simultaneously. European public health professionals operate across national, regional, and continental scales, with strong existing policy networks in addition to rich patterns of cross-border migration influencing epidemic dynamics. A majority of European countries also cooperate in setting policy with

inter-governmental European bodies such as the European Centre for Disease Prevention and Control (ECDC). In this case, a consistent approach to forecasting across the continent as a whole can support accurately informing cross-European monitoring, analysis, and guidance [3]. At a regional level, multi-country forecasts can support a better understanding of the impact of regional migration networks. Meanwhile, where there is limited capacity for infectious disease forecasting at a national level, forecasters generating multi-country results can provide an otherwise-unavailable opportunity for forecasts to inform national situational awareness. Some independent forecasting models have sought to address this by producing multi-country results [6]–[9].

Variation in forecast methods and presentation makes it difficult to compare predictive performance between forecast models, and from there to derive objective arguments for using one forecast over another. This confounds the selection of a single representative forecast and reduces the reliability of the evidence base for decisions based on forecasts. A "forecast hub" is a centralised effort to improve the transparency and usefulness of forecasts, by standardising and collating the work of many independent teams producing forecasts [10]. A hub sets a commonly agreed-upon structure for forecast targets, such as type of disease event, spatio-temporal units, or the set of quantiles of the probability distribution to include from probabilistic forecasts. For instance, a hub may collect predictions of the total number of cases reported in a given country for each day in the next two weeks. Forecasters can adopt this format and contribute forecasts for centralised storage in the public domain.

This shared infrastructure allows forecasts produced from diverse teams and methods to be visualised and quantitatively compared on a like-for-like basis, which can strengthen public and policy use of disease forecasts. The underlying approach to creating a forecast hub was pioneered in climate modelling and adapted for collaborative epidemiological forecasts of dengue [11] and influenza in the USA [10], [12]. This infrastructure was adapted for forecasts of short-term COVID-19 cases and deaths in the US [13], [14], prompting similar efforts in some European countries [15]–[17].

Standardising forecasts allows for combining multiple forecasts into a single ensemble with the potential for an improved predictive performance. Evidence from previous efforts in multi-model infectious disease forecasting suggests that forecasts from an ensemble of models can be consistently high performing compared to any one of the component models [11], [12], [18]. Elsewhere, weather forecasting has a long-standing use of building ensembles of models using diverse methods with standardised data and formatting in order to improve performance [19], [20].

The European COVID-19 Forecast Hub [21] is a project to collate short term forecasts of COVID-19 across 32 countries in the European region. The Hub is funded and supported by the ECDC, with the primary aim to provide reliable information about the near-term epidemiology of the COVID-19 pandemic to the

research and policy communities and the general public [3]. Second, the Hub aims to create infrastructure for storing and analysing epidemiological forecasts made in real time by diverse research teams and methods across Europe. Third, the Hub aims to maintain a community of infectious disease modellers underpinned by open science principles.

We started formally collating and combining contributions to the European Forecast Hub in March 2021. Here, we investigate the predictive performance of an ensemble of all forecasts contributed to the Hub in real time each week, as well as the performance of variations of ensemble methods created retrospectively.

Methods

We developed infrastructure to host and analyse prospective forecasts of COVID-19 cases and deaths. The infrastructure is compatible with equivalent research software from the US [22], [23] and German and Polish COVID-19 [24] Forecast Hubs, and easy to replicate for new forecasting collaborations.

Forecast targets and models

We sought forecasts for the incidence of COVID-19 as the total reported number of cases and deaths per week. We considered forecasts for 32 countries in Europe, including all countries of the European Union, European Free Trade Area, and the United Kingdom. We compared forecasts against observed data reported for each country by Johns Hopkins University (JHU, [25]). JHU data sources included a mix of national and aggregated subnational data. We aggregated incidence over the Morbidity and Mortality Weekly Report (MMWR) epidemiological week definition of Sunday through Saturday.

Teams could express their uncertainty around any single forecast target by submitting predictions for up to 23 quantiles (from 0.01 to 0.99) of the predictive probability distribution. Teams could also submit a single point forecast. At the first submission we asked teams to add a pre-specified set of metadata briefly describing the forecasting team and methods (see supplementary information (SI)). No restrictions were placed on who could submit forecasts. To increase participation we actively contacted known forecasting teams across Europe and the US and advertised among the ECDC network. Teams submitted a broad spectrum of model types, ranging from mechanistic to empirical models, agent-based and statistical models, and ensembles of multiple quantitative or qualitative models (described at [26]). We maintain a full project specification with a detailed submissions protocol [27].

We collected forecasts submitted weekly in real time over the 52 week period from 08 March 2021 to 07

March 2022. Teams submitted at latest two days after the complete dataset for the latest forecasting week became available each Sunday. We implemented an automated validation programme to check that each new forecast conformed to standardised formatting. Forecast validation ensured a monotonic increase of predictions with each increasing quantile, integer-valued non-negative counts of predicted cases, as well as consistent date and location definitions.

Each week we used all available valid forecasts to create a weekly real-time ensemble model (referred to as "the ensemble" from here on), for each of the 256 possible forecast targets: incident cases and deaths in 32 locations over the following one through four weeks. The ensemble method was an unweighted average of all models' forecast values, at each predictive quantile for a given location, target, and horizon. From 08 March 2021, we used the arithmetic mean. However we noticed that including highly anomalous forecasts in a mean ensemble produced extremely wide uncertainty. To mitigate this, from 26$^{\text{th}}$ July 2021 onwards the ensemble instead used a median of all predictive quantiles.

We created an open and publicly accessible interface to the forecasts and ensemble, including an online visualisation tool allowing viewers to see past data and interact with one or multiple forecasts for each country and target for up to four weeks' horizon [28]. All forecasts, metadata, and evaluations are freely available and held on Github [21] (archived in real-time at [29]), and Zoltar, a platform for hosting epidemiological forecasts [30], [31]. In the codebase for this study [32] we provide a simple method and instructions for downloading and preparing these data for analysis using R. We encourage other researchers to freely use and adapt this to support their own analyses.

Forecast evaluation

In this study we focused only on the comparative performance of forecasting models relative to each other. Performance in absolute terms is available on the Hub website [28]. For each model, we assessed calibration and overall predictive performance. We evaluated all previous forecasts against actual observed values for each model, stratified by the forecast horizon, location, and target. We calculated scores using the *scoringutils* R package [33]. We removed any forecast surrounding (both the week of, and the first week after) a strongly anomalous data point. We defined anomalous as where any subsequent data release revised that data point by over 5%.

To investigate calibration we assessed coverage as the correspondence between the forecast probability of an event and the observed frequency of that event. This usage follows previous work in epidemic forecasting [34], and is related to the concept of reliability for binary forecasts. We established the accuracy of each

model's prediction boundaries as the coverage of the predictive intervals. We calculated coverage at a given interval level k, where $k \in [0, 1]$, as the proportion p of observations that fell within the corresponding central predictive intervals across locations and forecast dates. A perfectly calibrated model would have $p = k$ at all 11 levels (corresponding to 22 quantiles excluding the median). An underconfident model at level k would have $p > k$, i.e. more observations fall within a given interval than expected. In contrast, an overconfident model at level k would have $p < k$, i.e. fewer observations fall within a given interval than expected. We here focus on coverage at the $k = 0.5$ and $k = 0.95$ levels.

We also assessed the overall predictive performance of weekly forecasts using the Weighted Interval Score~(WIS) across all available quantiles. The WIS represents a parsimonious approach to scoring forecasts based on uncertainty represented as forecast values across a set of quantiles [34], and is a strictly proper scoring rule, that is, it is optimal for predictions that come from the data-generating model. As a consequence, the WIS encourages forecasters to report predictions representing their true belief about the future [35]. Each forecast for a given location and date is scored based on an observed count of weekly incidence, the median of the predictive distribution and the predictive upper and lower quantiles corresponding to the central predictive interval level.

Not all models provided forecasts for all locations and dates, and we needed to compare predictive performance in the face of various levels of missingness across each forecast target. Therefore we calculated a relative WIS. This is a measure of forecast performance which takes into account that different teams may not cover the same set of forecast targets (i.e., weeks and locations). The relative WIS is computed using a *pairwise comparison tournament* where for each pair of models a mean score ratio is computed based on the set of shared targets. The relative WIS of a model with respect to another model is then the ratio of their respective geometric mean of the mean score ratios, such that smaller values indicate better performance.

We scaled the relative WIS of each model with the relative WIS of a baseline model, for each forecast target, location, date, and horizon. The baseline model assumes case or death counts stay the same as the latest data point over all future horizons, with expanding uncertainty, described previously in [36]. In this study we report the relative WIS of each model with respect to the baseline model.

Retrospective ensemble methods We retrospectively explored alternative methods for combining forecasts for each target at each week. A natural way to combine probability distributions available in the quantile format [37] used here is

$$F^{-1}(\alpha) = \sum_{i=1}^{n} w_i F_i^{-1}(\alpha),$$

Where $F_1 \dots F_n$ are the cumulative distribution functions of the individual probability distributions (in our case, the predictive distributions of each forecast model i contributed to the hub), w_i are a set of weights in $[0, 1]$; and α are the quantile levels, such that following notation introduced in [37],

$$F^{-1}(\alpha) = \inf\{t : F_i(t) \geq \alpha\}.$$

Different ensemble choices then mainly translate to the choice of weights w_i. An arithmetic mean ensemble uses weights at $w_i = 1/n$, where all weights are equal and sum up to 1.

Alternatively, we can choose a set of weights to apply to forecasts before they are combined. Numerous options exist for choosing these weights with the aim to maximise predictive performance, including choosing weights to reflect each forecast's past performance (thereby moving from an untrained to a trained ensemble). A straightforward choice is so-called inverse score weighting. In this case, the weights are calculated as

$$w_i = \frac{1}{S_i},$$

where S_i reflects the forecasting skill calculated as the relative WIS of forecaster i, calculated over all available model data, and normalised so that weights sum to 1. This method of weighting was found in the US to outperform unweighted scores during some time periods [38] but this was not confirmed in a similar study in Germany and Poland [15].

When constructing ensembles from quantile means, a single outlier can have an oversized effect on the ensemble forecast. Previous research has found that a median ensemble, replacing the arithmetic mean of each quantile with a median of the same values, yields competitive performance while maintaining robustness to outlying forecasts [39]. Building on this, we also created weighted median ensembles using the weights described above and a Harrel-Davis quantile estimator with a beta function to approximate the weighted percentiles [40]. We then compared the performance of unweighted and inverse relative WIS weighted mean and median ensembles.

Results

For 32 European countries, we collected, visualised, and made available online weekly COVID-19 forecasts and observed data [29]. An example of weekly forecasts from the ensemble model is shown in Figure 1.

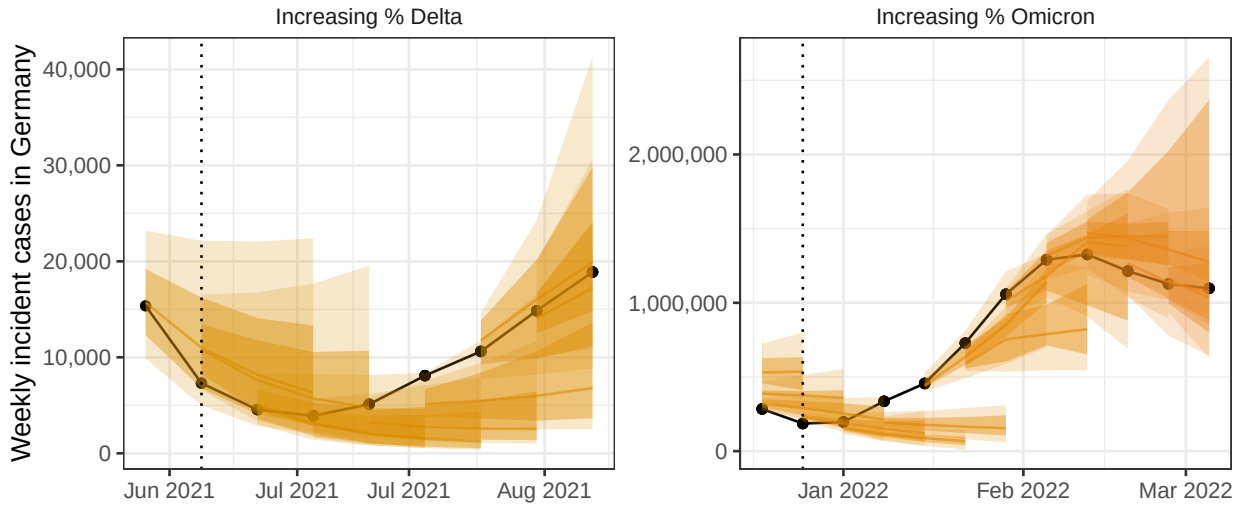

Figure 1: *Ensemble forecasts of weekly incident cases in Germany over periods of increasing SARS-CoV-2 variants Delta (B.1.617.2, left) and Omicron (B.1.1.529, right). Black indicates observed data. Coloured ribbons represent each weekly forecast of 1-4 weeks ahead (showing median, 50%, and 90% probability). For each variant, forecasts are shown over an x-axis bounded by the earliest dates at which 5% and 99% of sequenced cases were identified as the respective variant of concern, while vertical dotted lines indicate the approximate date that the variant reached dominance (>50% sequenced cases).*

Over the whole study period, we collected forecasts from 48 unique models. Modellers created forecasts choosing from a set of 32 possible locations, four time horizons, and two variables, and modellers variously joined and left the Hub over time. This meant the number of models contributing to the Hub varied over time and by forecasting target. Of the total 48 models, we received the most forecasts for Germany, with 29 unique models submitting one-week case forecasts, while only 12 models ever submitted four-week case or death forecasts for Liechtenstein. Modelling teams also differed in how they expressed uncertainty. Only 3 models provided point forecasts with no estimate of uncertainty around their predictions, while 41 models provided the full set of 23 probabilistic quantiles across the predictive distribution for each target.

In this evaluation we included 29 models in comparison to the ensemble forecast (SI Figure 1). We have included metadata provided by modellers in the supplement (SI Table 1), and also online [29]. At most, 15 models contributed forecasts for cases in Germany at the 1 week horizon, with an accumulated 592 forecast scores for that single target over the study period. In contrast, deaths in Finland at the 2 week horizon saw the smallest number of forecasts, with only 6 independent models contributing 24 forecast scores at any time

over the 52 week period. Of the 29 models included in this evaluation, 5 models provided less than the full set of 23 quantiles, and were excluded when creating the ensemble. No ensemble forecast was composed of less than 3 independent models.

Using all models and the ensemble, we created 2139 forecasting scores where each score summarises a unique combination of forecasting model, variable, country, and week ahead horizon (SI Figure 2). We visually compared the absolute performance of forecasts in predicting numbers of incident cases and deaths. We observed that forecasts predicted well in times of stable epidemic behaviour, while struggling to accurately predict at longer horizons around inflection points, for example during rapid changes in population-level behaviour or surveillance. Forecast models varied widely in their ability to predict and account for the introduction of new variants, giving the ensemble forecast over these periods a high level of uncertainty (Figure 1).

In relative terms, the ensemble of all models performed well compared to both its component models and the baseline. By relative WIS scaled against a baseline of 1 (where a score <1 indicates outperforming the baseline), the median score of forecasts from the Hub ensemble model was 0.71, within an interquartile range of 0.61 at 25% probability to 0.88 at 75% probability. Meanwhile the median score of forecasts across all participating models (excluding the Hub ensemble) was 1.04 (IQR 0.82-1.36).

Across all horizons and locations, the ensemble performed better on scaled relative WIS than 83% of forecast scores when forecasting cases (with a total N=886 from 23 unique models), and 91% of scores for forecasts of incident deaths (N=763 scores from 20 models). We also saw high performance from the ensemble when evaluating against all models including those who did not submit the full set of probabilistic quantile predictions (80% for cases with N=1006 scores from 28 models, and 88% for deaths, N=877 scores from 24 models).

The performance of individual and ensemble forecasts varied by length of the forecast horizon (Figure 2). At each horizon, the typical performance of the ensemble outperformed both the baseline model and the aggregated scores of all its component models, although we saw wide variation between individual models in performance across horizons. Both individual models and the ensemble saw a trend of worsening performance at longer horizons when forecasting cases with the median scaled relative WIS of the ensemble across locations worsened from 0.62 for one-week ahead forecasts to 0.9 when forecasting four weeks ahead. Performance for forecasts of deaths was more stable over one through four weeks, with median ensemble performance moving from 0.69 to 0.76 across the four week horizons.

We observed similar trends in performance across horizon when considering how well the ensemble was

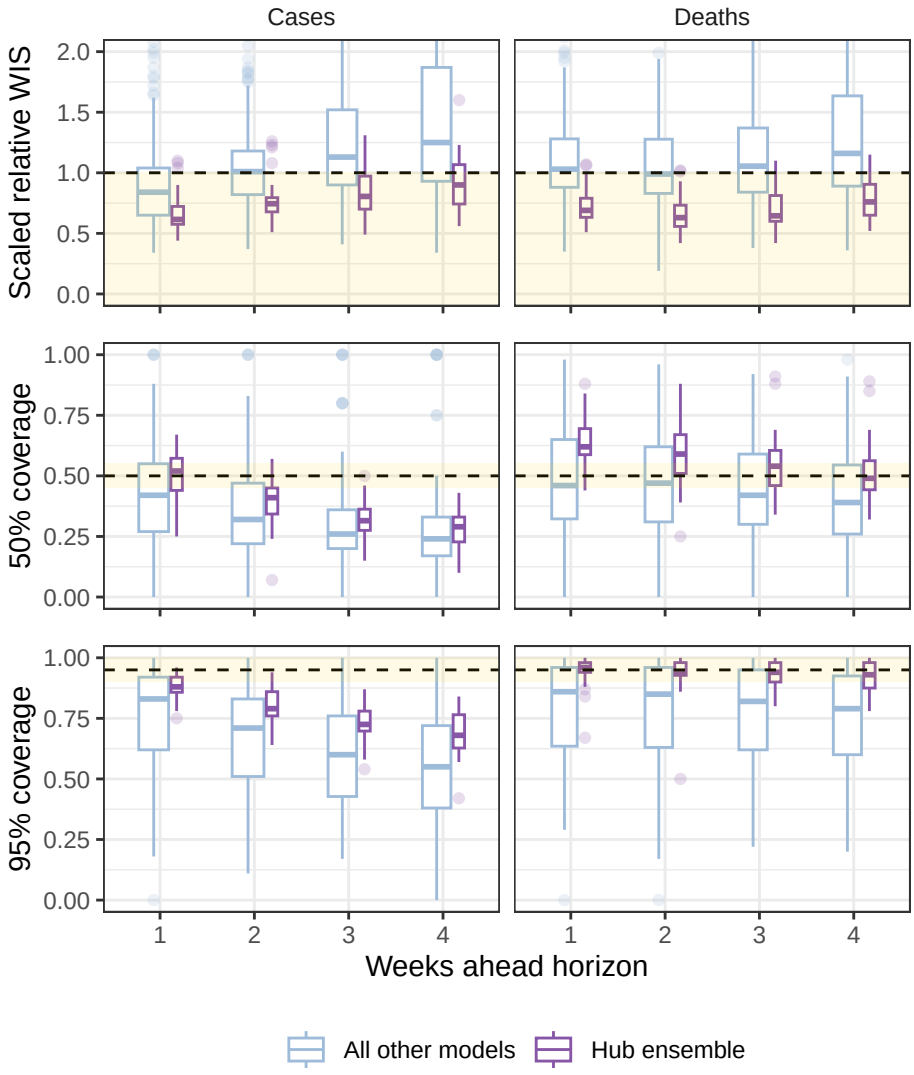

Figure 2: *Performance of short-term forecasts aggregated across all individually submitted models and the Hub ensemble, by horizon, forecasting cases (left) and deaths (right). Performance measured by relative weighted interval score scaled against a baseline (dotted line, 1), and coverage of uncertainty at the 50% and 95% levels. Boxplot, with width proportional to number of observations, show interquartile ranges with outlying scores as faded points. The target range for each set of scores is shaded in yellow.*

calibrated with respect to the observed data. At one week ahead the case ensemble was well calibrated (ca. 50% and 95% nominal coverage at the 50% and 95% levels respectively). This did not hold at longer forecast horizons as the case forecasts became increasingly over-confident. Meanwhile, the ensemble of death forecasts was well calibrated at the 95% level across all horizons, and the calibration of death forecasts at the 50% level improved with lengthening horizons compared to being underconfident at shorter horizons.

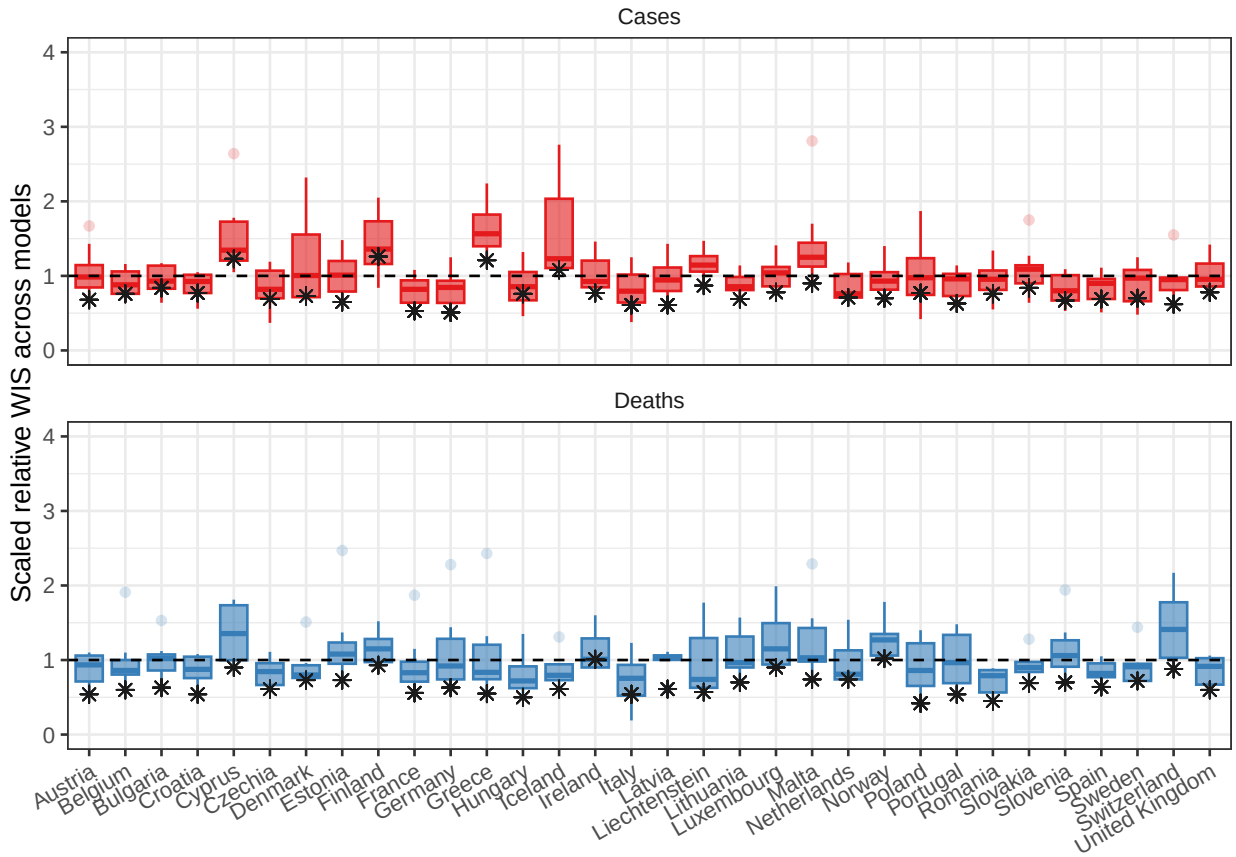

Figure 3: *Performance of short-term forecasts across models and median ensemble (asterisk), by country, forecasting cases (top) and deaths (bottom) for two-week ahead forecasts, according to the relative weighted interval score. Boxplots show interquartile ranges, with outliers as faded points, and the ensemble model performance is marked by an asterisk. y-axis is cut-off to an upper bound of 4 for readability.*

The ensemble also performed consistently well in comparison to individual models when forecasting across countries (Figure 3). In total, across 32 countries forecasting for one through four weeks, when forecasting cases the ensemble outperformed 75% of component models in 22 countries, and outperformed all available models in 3 countries. When forecasting deaths, the ensemble outperformed 75% and 100% of models in 30 and 8 countries respectively. Considering only the the two-week horizon shown in Figure 3, the ensemble of case forecasts outperformed 75% models in 25 countries and all models in only 12 countries. At the two-week

Table 1: Predictive performance of main ensembles, as measured by the scaled relative WIS.

Horizon	Weighted mean	Weighted median	Unweighted mean	Unweighted median
Cases				
1 week	0.59	0.62	0.59	0.61
2 weeks	0.67	0.67	0.67	0.67
3 weeks	0.79	0.70	0.81	0.71
4 weeks	1.06	0.75	1.09	0.79
Deaths				
1 week	0.63	0.59	1.00	0.59
2 weeks	0.57	0.54	0.81	0.53
3 weeks	0.64	0.56	0.83	0.54
4 weeks	0.83	0.64	0.82	0.62

horizon for forecasts of deaths, the ensemble outperformed 75% and 100% of its component models in 30 and 26 countries respectively.

We considered alternative methods for creating ensembles from the participating forecasts, using either a mean or median to combine either weighted or unweighted forecasts (Table 1). Across locations we observed that the median outperformed the mean across all one through four week horizons and both cases and death targets, for all but cases at the 1 week horizon. This held regardless of whether the component forecasts were weighted or unweighted by their individual past performance. Between methods of combination, weighting made little difference to the performance of the median ensemble, but appeared to improve performance of a mean ensemble in forecasting deaths.

Discussion

We collated 12 months of forecasts of COVID-19 cases and deaths across 32 countries in Europe, collecting from multiple independent teams and using a principled approach to standardising both forecast targets and the predictive distribution of forecasts. We combined these into an ensemble forecast and compared the relative performance of forecasts between models, finding that the ensemble forecasts outperformed most individual models across all countries and horizons over time.

Across all models we observed that forecasting changes in trend in real time was particularly challenging. Our study period included multiple fundamental changes in viral-, individual-, and population-level factors driving the transmission of COVID-19 across Europe. In early 2021, the introduction of vaccination started to change population-level associations between infections, cases, and deaths [41], while the Delta variant emerged and became dominant [42]. Similarly from late 2021 we saw the interaction of individually waning

immunity during the emergence and global spread of the Omicron variant [43]. Neither the extent nor timing of these factors were uniform across European countries covered by the Forecast Hub [44]. This meant that the performance of any single forecasting model depended partly on the ability, speed, and precision with which it could adapt to new conditions for each forecast target.

We observed a contrast between a more stable performance of forecasting deaths further into the future compared to forecasts of cases. Previous work has found rapidly declining performance for case forecasts with increasing horizon [36], [45], while death forecasts can perform well with up to six weeks lead time [46]. We can link this to the specific epidemic dynamics in this study.

First, COVID-19 has a typical serial interval of less than a week [47]. This implies that case forecasts of more than two weeks only remain valid if rates of both transmission and detection remain stable over the entire forecast horizon. In contrast, we saw rapid changes in epidemic dynamics across many countries in Europe over our study period, impacting the longer term case forecasts.

Second, we can interpret the higher reliability of death forecasts as due to the different lengths and distributions of time lags from infection to case and death reporting [48]. For example, a spike in infections may be matched by a consistently sharp increase in case reporting, but a longer-tailed distribution of the subsequent increase in death reports. This creates a lower magnitude of fluctuation in the time-series of deaths compared to that of cases. Similarly, surveillance data for death reporting is substantially more consistent, with fewer errors and retrospective corrections, than case reporting [49].

Third, we also note that the performance of trend-based forecasts may have benefited from the slower changes to trends in incident deaths caused by gradually increasing vaccination rates. These features allow forecasters to incorporate the effect of changes in transmission more easily when forecasting deaths, compared to cases.

We found the ensemble in this study continued to outperform both other models and the baseline at up to four weeks ahead. Our results support previous findings that ensemble forecasts are the best or nearly the best performing models with respect to absolute predictive performance and appropriate coverage of uncertainty [16], [18], [36]. While the ensemble was consistently high performing, it was not strictly dominant across all forecast targets, reflecting findings from previous comparable studies of COVID-19 forecasts [15], [50]. Our finding suggests the usefulness of an ensemble as a robust summary when forecasting across many spatio-temporal targets, without replacing the importance of communicating the full range of model predictions.

When exploring variations in ensemble methods, we found that the choice of median over means yielded the most consistent improvement in predictive performance, regardless of the method of weighting. Other

work has supported the importance of the median in providing a stable forecast that better accounts for outlier forecasts than the mean [50], although this finding may be dependent on the quality of the individual forecast submissions. In contrast, weighing models by past performance did not result in any consistent improvement in performance. This is in line with existing mixed evidence for any optimal ensemble method for combining short term probabilistic infectious disease forecasts. Many methods of combination have performed competitively in analyses of forecasts for COVID-19 in the US, including the simple mean and weighted approaches outperforming unweighted or median methods [38]. This contrasts with later analyses finding weighted methods to give similar performance to a median average [14], [50]. We can partly explain this inconsistency if performance of each method depends on the outcome being predicted (cases, deaths), its count (incident, cumulative) and absolute level, the changing disease dynamics, and the varying quality and quantity of forecasting teams over time.

We note several limitations in our approach to assessing the relative performance of an ensemble among forecast models. While we have described differences in model scores, we have not used any formal statistical test for comparing forecast scores, such as the Diebold-Mariano test [51], recognising that it is unclear how this is best achieved across many models. Our results are the outcome of evaluating forecasts against a specific performance metric and baseline, where multiple options for evaluation exist and the choice reflects the aim of the evaluation process. Further, our choice of baseline model affects the given performance scores in absolute terms, and more generally the choice of appropriate baseline for epidemic forecast models is not obvious when assessing infectious disease forecasts. The model used here is supported by previous work [36], yet previous evaluation in a similar context has suggested that choice of baseline affects relative performance in general [52], and future research should be done on the best choices of baseline models in the context of infectious disease epidemics.

Our assessment of forecast performance may further have been inaccurate due to limitations in the observed data against which we evaluated forecasts. We sourced data from a globally aggregated database to maintain compatibility across 32 countries [25]. However, this made it difficult to identify the origin of lags and inconsistencies between national data streams, and to what extent these could bias forecasts for different targets. In particular we saw some real time data revised retrospectively, introducing bias in either direction where the data used to create forecasts was not the same as that used to evaluate it. We attempted to mitigate this by using an automated process for determining data revisions, and excluding forecasts made at a time of missing, unreliable, or heavily revised data. We also recognise that evaluating forecasts against updated data is a valid alternative approach used elsewhere [36]. More generally it is unclear if the expectation of observation revisions should be a feature built into forecasts. Further research is needed to understand the

perspective of end-users of forecasts in order to assess this.

The focus of this study was describing and summarising an ensemble of many models. We note that we have little insight into the individual methods and wide variety of assumptions that modellers used. While we asked modellers to provide a short description of their methods, we did not create a rigorous framework for this, and we did not document whether modellers changed the methods for a particular submitted model over time. Both the content of and variation in modelling methods and assumptions are likely to be critical to explaining performance, rather than describing or summarising it. Exploring modellers' methods and relating this to forecast performance will be an important area of future work.

In an emergency setting, access to visualised forecasts and underlying data is useful for researchers, policy-makers, and the public [2]. Previous European multi-country efforts to forecast COVID-19 have included only single models adapted to country-specific parameters [6], [7], [9].

The European Forecasting Hub acted as a unique tool for creating an open-access, cross-country modelling network, and connecting this to public health policy across Europe. By opening participation to many modelling teams and with international high participation, we were able to create robust ensemble forecasts across Europe. This also allows comparison across forecasts built with different interpretations of current data, on a like for like scale in real time. The European Hub has supported policy outputs at an international, regional, and national level, including Hub forecasts cited weekly in ECDC Communicable Disease Threats Reports ([53]).

For forecast producers, an easily accessible comparison between results from different methods can highlight individual strengths and weaknesses and help prioritise new areas of work. Collating time-stamped predictions ensures that we can test true out-of-sample performance of models and avoid retrospective claims of performance. Testing the limits of forecasting ability with these comparisons forms an important part of communicating any model-based prediction to decision makers. For example, the weekly ECDC Communicable Disease Threats reports include the specific results of this work by qualitatively highlighting the greater uncertainty around case forecasts compared to death forecasts.

This study raises many further questions which could inform epidemic forecast modellers and users. The dataset created by the European Forecast Hub is an openly accessible, standardised, and extensively documented catalogue of real time forecasting work from a range of teams and models across Europe [28], and we recommend its use for further research on forecast performance. In the code developed for this study we provide a worked example of downloading and using both the forecasts and their evaluation scores [32].

Future work could explore the impact on forecast models of changing epidemiology at a broad spatial scale by

combining analyses of trends and turning points in cases and deaths with forecast performance, or extending to include data on vaccination, variant, or policy changes over time. There is also much scope for future research into methods for combining forecasts to improve performance of an ensemble. This includes altering the inclusion criteria of forecast models based on different thresholds of past performance, excluding or including only forecasts that predict the lowest and highest values (trimming) [38], or using alternative weighting methods such as quantile regression averaging [16]. Exploring these questions would add to our understanding of real time performance, supporting and improving future forecasting efforts.

We see additional scope to adapt the Hub format to the changing COVID-19 situation across Europe. We have extended the Forecast Hub infrastructure to include short term forecasts for hospitalisations with COVID-19, which is a challenging task due to limited data across the locations covered by the hub. As the policy focus shifts from immediate response to anticipating changes brought by vaccinations or the geographic spread of new variants [44], we are also separately investigating models for longer term scenarios in addition to the short term forecasts in a similar framework to existing scenario modelling work in the US [54].

In conclusion, we have shown that during a rapidly evolving epidemic spreading through multiple populations, an ensemble forecast performed highly consistently across a large matrix of forecast targets, typically outperforming the majority of its separate component models and a naive baseline model. In addition, we have linked issues with the predictability of short-term case forecasts to underlying COVID-19 epidemiology, and shown that ensemble methods based on past model performance were unable to reliably improve forecast performance. Our work constitutes a step towards both unifying COVID-19 forecasts and improving our understanding of them.

References

[1] P. van Basshuysen, L. White, D. Khosrowi, and M. Frisch, "Three Ways in Which Pandemic Models May Perform a Pandemic," *Erasmus Journal for Philosophy and Economics*, vol. 14, no. 1, 1, pp. 110-127-110-127, Jul. 2021, doi: 10.23941/ejpe.v14i1.582.

[2] CDC, "Coronavirus Disease 2019 (COVID-19)," Feb. 11, 2020. https://www.cdc.gov/coronavirus/2019-ncov/science/forecasting/forecasting.html (accessed Jan. 09, 2022).

[3] European Centre for Disease Prevention and Control, "Forecasting COVID-19 cases and deaths in Europe - new hub will support European pandemic planning," Apr. 22, 2021. https://www.ecdc.europa.eu/en/news-events/forecasting-covid-19-cases-and-deaths-europe-new-hub

[4] J. Zelner, J. Riou, R. Etzioni, and A. Gelman, "Accounting for uncertainty during a pandemic," *PATTER*, vol. 2, no. 8, Aug. 2021, doi: 10.1016/j.patter.2021.100310.

[5] L. P. James, J. A. Salomon, C. O. Buckee, and N. A. Menzies, "The Use and Misuse of Mathematical Modeling for Infectious Disease Policymaking: Lessons for the COVID-19 Pandemic," *Med Decis Making*, vol. 41, no. 4, pp. 379–385, May 2021, doi: 10.1177/0272989X21990391.

[6] R. Aguas *et al.*, "Modelling the COVID-19 pandemic in context: An international participatory approach," *BMJ Global Health*, vol. 5, no. 12, p. e003126, Dec. 2020, doi: 10.1136/bmjgh-2020-003126.

[7] K. Adib *et al.*, "A participatory modelling approach for investigating the spread of COVID-19 in countries of the Eastern Mediterranean Region to support public health decision-making," *BMJ Global Health*, vol. 6, no. 3, p. e005207, Mar. 2021, doi: 10.1136/bmjgh-2021-005207.

[8] A. Agosto and P. Giudici, "A poisson autoregressive model to understand COVID-19 contagion dynamics," *Risks*, vol. 8, no. 3, p. 77, Jul. 2020, doi: 10.3390/risks8030077.

[9] A. Agosto, A. Campmas, P. Giudici, and A. Renda, "Monitoring COVID-19 contagion growth," *Statistics in Medicine*, vol. 40, no. 18, pp. 4150–4160, 2021, doi: 10.1002/sim.9020.

[10] N. G. Reich *et al.*, "A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States," *PNAS*, vol. 116, no. 8, pp. 3146–3154, Feb. 2019, doi: 10.1073/pnas.1812594116.

[11] M. A. Johansson *et al.*, "An open challenge to advance probabilistic forecasting for dengue epidemics," *PNAS*, vol. 116, no. 48, pp. 24268–24274, Nov. 2019, doi: 10.1073/pnas.1909865116.

[12] N. G. Reich *et al.*, "Accuracy of real-time multi-model ensemble forecasts for seasonal influenza in the U.S," *PLoS Comput Biol*, vol. 15, no. 11, p. e1007486, Nov. 2019, doi: 10.1371/journal.pcbi.1007486.

[13] E. Y. Cramer *et al.*, "The United States COVID-19 Forecast Hub dataset," p. 2021.11.04.21265886, Nov. 2021, doi: 10.1101/2021.11.04.21265886.

[14] E. L. Ray *et al.*, "Ensemble Forecasts of Coronavirus Disease 2019 (COVID-19) in the U.S." p. 2020.08.19.20177493, Aug. 2020, doi: 10.1101/2020.08.19.20177493.

[15] J. Bracher *et al.*, "A pre-registered short-term forecasting study of COVID-19 in Germany and Poland during the second wave," *Nat Commun*, vol. 12, no. 1, 1, p. 5173, Aug. 2021, doi: 10.1038/s41467-021-25207-0.

[16] S. Funk *et al.*, "Short-term forecasts to inform the response to the Covid-19 epidemic in the UK," *medRxiv*, p. 2020.11.11.20220962, Nov. 2020, doi: 10.1101/2020.11.11.20220962.

[17] M. Bicher *et al.*, "Supporting COVID-19 Policy-Making with a Predictive Epidemiological Multi-Model Warning System," *medRxiv*, p. 2020.10.18.20214767, Apr. 2021, doi: 10.1101/2020.10.18.20214767.

[18] C. Viboud *et al.*, "The RAPIDD ebola forecasting challenge: Synthesis and lessons learnt," *Epidemics*, vol. 22, pp. 13–21, Mar. 2018, doi: 10.1016/j.epidem.2017.08.002.

[19] R. Buizza, "Introduction to the special issue on '25 years of ensemble forecasting'," *Quarterly Journal of the Royal Meteorological Society*, vol. 145, no. S1, pp. 1–11, 2019, doi: 10.1002/qj.3370.

[20] K. R. Moran *et al.*, "Epidemic Forecasting is Messier Than Weather Forecasting: The Role of Human Behavior and Internet Data Streams in Epidemic Forecast," *J Infect Dis*, vol. 214, pp. S404–S408, Dec. 2016, doi: 10.1093/infdis/jiw375.

[21] European Covid-19 Forecast Hub, *European COVID-19 Forecast Hub*. covid19-forecast-hub-europe, 2021.Available: https://github.com/covid19-forecast-hub-europe/covid19-forecast-hub-europe

[22] E. Cramer *et al.*, "Reichlab/Covid19-forecast-hub: Release for Zenodo, 20210816," Aug. 2021, doi: 10.5281/zenodo.5208210.

[23] S. Y. Wang *et al.*, "Reichlab/covidHubUtils: Repository release for Zenodo," Aug. 2021, doi: 10.5281/zenodo.5207940.

[24] J. Bracher *et al.*, *The German and Polish COVID-19 Forecast Hub*. 2020.Available: https://github.com/KITmetricslab/covid19-forecast-hub-de

[25] E. Dong, H. Du, and L. Gardner, "An interactive web-based dashboard to track COVID-19 in real time," *The Lancet Infectious Diseases*, vol. 20, no. 5, pp. 533–534, May 2020, doi: 10.1016/S1473-3099(20)30120-1.

[26] European Covid-19 Forecast Hub, "Community." https://covid19forecasthub.eu/community.html

[27] European Covid-19 Forecast Hub, "Covid19-forecast-hub-europe: Wiki." https://github.com/covid19-forecast-hub-europe/covid19-forecast-hub-europe

[28] European Covid-19 Forecast Hub, "European Covid-19 Forecast Hub." https://covid19forecasthub.eu/index.html

[29] K. Sherratt *et al.*, "European covid-19 forecast hub." Zenodo, Nov. 2022. doi: 10.5281/zenodo.7356267.

[30] EpiForecasts, "Project: ECDC European COVID-19 Forecast Hub - Zoltar," 2021. https://www.zoltardata.com/project/238

[31] N. G. Reich, M. Cornell, E. L. Ray, K. House, and K. Le, "The Zoltar forecast archive, a tool to standardize and store interdisciplinary prediction research," *Sci Data*, vol. 8, no. 1, 1, p. 59, Feb. 2021, doi: 10.1038/s41597-021-00839-5.

[32] *Predictive performance of multi-model ensemble forecasts of Covid-19 across European nations*. covid19-forecast-hub-europe, 2022.Available: https://github.com/covid19-forecast-hub-europe/euro-hub-ensemble

[33] Nikos I Bosse, Hugo Gruson, Sebastian Funk, EpiForecasts, and Sam Abbott, *Scoringutils: Utilities for Scoring and Assessing Predictions*. 2020.Available: https://github.com/epiforecasts/scoringutils

[34] J. Bracher, E. L. Ray, T. Gneiting, and N. G. Reich, "Evaluating epidemic forecasts in an interval format," *PLOS Computational Biology*, vol. 17, no. 2, p. e1008618, Feb. 2021, doi: 10.1371/journal.pcbi.1008618.

[35] T. Gneiting and A. E. Raftery, "Strictly Proper Scoring Rules, Prediction, and Estimation," *Journal of the American Statistical Association*, vol. 102, no. 477, pp. 359–378, Mar. 2007, doi: 10.1198/016214506000001437.

[36] E. Y. Cramer *et al.*, "Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the US," *medRxiv*, p. 2021.02.03.21250974, Jan. 2021, doi: 10.1101/2021.02.03.21250974.

[37] C. Genest, "Vincentization Revisited," *The Annals of Statistics*, vol. 20, no. 2, pp. 1137–1142, 1992,Available: https://www.jstor.org/stable/2242003

[38] J. W. Taylor and K. S. Taylor, "Combining Probabilistic Forecasts of COVID-19 Mortality in the United States," *Eur J Oper Res*, Jun. 2021, doi: 10.1016/j.ejor.2021.06.044.

[39] E. L. Ray *et al.*, "Comparing trained and untrained probabilistic ensemble forecasts of COVID-19 cases and deaths in the United States," Jan. 28, 2022. Accessed: Mar. 30, 2022. [Online]. Available: http://arxiv.org/abs/2201.12387

[40] F. E. HARRELL and C. E. DAVIS, "A new distribution-free quantile estimator," *Biometrika*, vol. 69, no. 3, pp. 635–640, Dec. 1982, doi: 10.1093/biomet/69.3.635.

[41] European Centre for Disease Prevention and Control, "Interim guidance on the benefits of full vaccination against COVID-19 for transmission and implications for non-pharmaceutical interventions - 21 April 2021," ECDC, Stockholm, 2021.Available: https://www.ecdc.europa.eu/en/publications-data/interim-guidance-benefits-full-vaccination-against-covid-19-transmission

[42] European Centre for Disease Prevention and Control, "Threat Assessment Brief: Implications for the EU/EEA on the spread of the SARS-CoV-2 Delta (B.1.617.2) variant of concern," ECDC, Stockholm, Jun. 2021.Available: https://www.ecdc.europa.eu/en/publications-data/threat-assessment-emergence-and-impact-sars-cov-2-delta-variant

[43] European Centre for Disease Prevention and Control, "Assessment of the further spread and potential impact of the SARS-CoV-2 Omicron variant of concern in the EU/EEA, 19th update," Jan. 27, 2022. https://www.ecdc.europa.eu/en/publications-data/covid-19-omicron-risk-assessment-further-emergence-and-potential-impact

[44] European Centre for Disease Prevention and Control, "Overview of the implementation of COVID-19 vaccination strategies and deployment plans in the EU/EEA," ECDC, Stockholm, Nov. 2021.Available: https://www.ecdc.europa.eu/en/publications-data/overview-implementation-covid-19-vaccination-strategies-and-deployment-plans

[45] M. Castro, S. Ares, J. A. Cuesta, and S. Manrubia, "The turning point and end of an expanding epidemic cannot be precisely forecast," *Proceedings of the National Academy of Sciences*, vol. 117, no. 42, pp. 26190–26196, Oct. 2020, doi: 10.1073/pnas.2007868117.

[46] J. Friedman *et al.*, "Predictive performance of international COVID-19 mortality forecasting models," *Nat Commun*, vol. 12, no. 1, 1, p. 2609, May 2021, doi: 10.1038/s41467-021-22457-w.

[47] M. Alene, L. Yismaw, M. A. Assemie, D. B. Ketema, W. Gietaneh, and T. Y. Birhan, "Serial interval and incubation period of COVID-19: A systematic review and meta-analysis," *BMC Infectious Diseases*, vol. 21, no. 1, p. 257, Mar. 2021, doi: 10.1186/s12879-021-05950-x.

[48] R. Jin, "The lag between daily reported Covid-19 cases and deaths and its relationship to age," *J Public Health Res*, vol. 10, no. 3, p. 2049, Mar. 2021, doi: 10.4081/jphr.2021.2049.

[49] M. Català *et al.*, "Robust estimation of diagnostic rate and real incidence of COVID-19 for European policymakers," *PLOS ONE*, vol. 16, no. 1, p. e0243701, Jan. 2021, doi: 10.1371/journal.pone.0243701.

[50] L. Brooks, "Comparing ensemble approaches for short-term probabilistic COVID-19 forecasts in the U.S." 2020. https://forecasters.org/blog/2020/10/28/comparing-ensemble-approaches-for-short-term-probabilistic-covid-19-forecasts-in-the-u-s/ (accessed Jul. 15, 2021).

[51] F. X. Diebold and R. S. Mariano, "Comparing Predictive Accuracy," *Journal of Business & Economic Statistics*, vol. 13, no. 3, pp. 253–263, Jul. 1995, doi: 10.1080/07350015.1995.10524599.

[52] J. Bracher *et al.*, "National and subnational short-term forecasting of COVID-19 in Germany and Poland, early 2021," p. 2021.11.05.21265810, Nov. 2021, doi: 10.1101/2021.11.05.21265810.

[53] European Centre for Disease Prevention and Control, "Weekly threats reports (CDTR)," 2022. https://www.ecdc.europa.eu/en/publications-and-data/monitoring/weekly-threats-reports

[54] R. K. Borchering, "Modeling of Future COVID-19 Cases, Hospitalizations, and Deaths, by Vaccination Rates and Nonpharmaceutical Intervention Scenarios — United States, April–September 2021," *MMWR Morb Mortal Wkly Rep*, vol. 70, 2021, doi: 10.15585/mmwr.mm7019e3.