

¹ Comparing human and model-based forecasts of COVID-19 in
² Germany and Poland

³

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

⁷ **1 Abstract**

⁸ Forecasts based on epidemiological modelling have played an important role in shaping public policy throughout
⁹ the COVID-19 pandemic. This modelling combines knowledge about infectious disease dynamics with the
¹⁰ subjective opinion of the researcher who develops and refines the model and often also adjusts model outputs.
¹¹ Developing a forecast model is difficult, resource- and time-consuming. It is therefore worth asking what
¹² modelling is able to add beyond the subjective opinion of the researcher alone. To investigate this, we
¹³ analysed different real-time forecasts of cases of and deaths from COVID-19 in Germany and Poland over a
¹⁴ 1-4 week horizon submitted to the German and Polish Forecast Hub. We compared crowd forecasts elicited
¹⁵ from researchers and volunteers, against a) forecasts from two semi-mechanistic models based on common
¹⁶ epidemiological assumptions and b) the ensemble of all other models submitted to the Forecast Hub. We found
¹⁷ crowd forecasts, despite being overconfident, to outperform all other methods across all forecast horizons when
¹⁸ forecasting cases (weighted interval score relative to the Hub ensemble 2 weeks ahead: 0.89). Forecasts based
¹⁹ on computational models performed comparably better when predicting deaths (rel. WIS 1.26), suggesting
²⁰ that epidemiological modelling and human judgement can complement each other in important ways.

21 2 Author summary

22 Mathematical models of COVID-19 have played a key role in informing governments across the world. While
23 mathematical models are informed by our knowledge of infectious disease dynamics, they are ultimately
24 developed and iteratively adjusted by the researchers and shaped by their subjective opinions. To investigate
25 what modelling is able to add beyond the subjective opinion of the researcher alone, we compared human fore-
26 casts with model-based predictions of COVID-19 cases and deaths submitted to the so-called German/Polish
27 Forecast Hub (which collates a variety of models from a range of teams).

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

35 3 Introduction

36 Infectious disease modelling has a long tradition and has helped inform public health decisions both through
37 scenario modelling, as well as actual forecasts of (among others) influenza (e.g. 1,2–4), dengue fever (e.g.
38 5,6,7), ebola (e.g. 8,9), chikungunya (e.g. 10,11) and now COVID-19 (e.g. 12,13–17). Applications of
39 epidemiological models differ in the way they make statements about the future. Forecasts aim to predict the
40 future as it will occur, while scenario modelling and projections aim to represent what the future could look
41 like under certain scenario assumptions or if conditions stayed the same as they were in the past. Forecasts
42 can be judged by comparing them against observed data. Since it is much harder to fairly assess the accuracy
43 and usefulness of projections and scenario modelling in the same way, this work focuses on forecasts, which
44 represent only a subset of all epidemiological modelling.

45 Since March 2020, forecasts of COVID-19 from multiple teams have been collected, aggregated and compared

46 by Forecast Hubs such as the US Forecast Hub (13,14), the German and Polish Forecast Hub (15,16) and the
47 European Forecast Hub (17). Often, different individual forecasts are combined into a single forecast, e.g. by
48 taking the mean or median of all forecasts. These ensemble forecasts usually tend to perform better and
49 more consistently than individual forecasts (see e.g. (6); (18)).

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

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

65 Previous work has examined such direct human forecasts in various contexts, such as geopolitics (19,20),
66 meta-science (21,22), sports (23) and epidemiology (11,24,25). Several prediction platforms (26–28) and
67 prediction markets (29) have been created to collate expert and non-expert predictions. However, with the
68 notable exception of (11), these forecasts were not designed to be evaluated alongside model-based forecasts
69 and usually follow their own (often binary) prediction formats. Direct human forecasts may be able to take
70 into account insights and relationships between variables which are hard to specify using epidemiological
71 models. However, it is not entirely clear in which situations human forecasts perform well or badly. For

72 example, (11) found that humans could outperform computer models at predicting the 2014/15 and 2015/16
73 flu season in the US, a setting where the disease was well known and information about previous seasons was
74 available. However, humans tended to do slightly worse at predicting the 2014/15 outbreak of chikungunya
75 in the Americas, a disease previously largely unobserved and unknown in these regions at the time.

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

83 4 Methods

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

97 **4.1 Forecast targets and interaction with the German and Polish Forecast Hub**

98 The German and Polish Forecast Hub (now mostly merged into the (17)) elicits predictions for various
99 COVID-19 related forecast targets from different research groups every week. Forecasts had to be made
100 every Monday (with submissions allowed until Tuesday 3pm) and were permitted to use any data that was
101 available by Monday 11.59pm. We submitted forecasts for incident and cumulative weekly reported numbers
102 of cases of and deaths from COVID-19 on a national level in Germany and Poland over a one to four week
103 forecast horizon. Forecasts were submitted on Mondays, but weeks were defined as ending on a Saturday
104 (and starting on Sunday), meaning that forecast horizons were in fact 5, 12, 19 and 26 days. Submissions
105 were required in a quantile-based format with 23 quantiles of each output measure at levels 0.01, 0.025, 0.05,
106 0.10, 0.15, . . . , 0.95, 0.975, 0.99. Forecasts submitted to the Forecast Hub were combined into different
107 ensembles every week, with the median ensemble (i.e., the α -quantile of the ensemble is given by the median
108 of all submitted α -quantiles) being the default ensemble shown on all official Forecast hub visualisations
109 (<https://kitmetricslab.github.io/forecasthub/forecast>).

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

119 **4.2 Crowd forecasts**

120 Our crowd forecasts were created as an ensemble of forecasts made by individual participants every week
121 through a web application (<https://cmmid-lshtm.shinyapps.io/crowd-forecast/>). Weekly forecasts had to
122 be submitted before Tuesday 12pm every week, but participants were asked to only use any information or

123 data that was already available by Monday night. The application was built using the `shiny` and `golem`
124 R packages (37,38) and is available in the `crowdforecastr` R package (31). To make a forecast in the
125 application participants could select a predictive distribution (with the default being log-normal) to represent
126 the probability that the forecasted quantity took certain values. Median and width of the uncertainty could
127 be adjusted by either interacting with a figure showing their forecast or providing numerical values (see
128 screenshot in Figure S1 in the SI). The default shown was a repetition of the last known observation with
129 constant uncertainty around it computed as the standard deviation of the last four changes in weekly log
130 observed forecasts (i.e. as $\sigma(\log(value4) - \log(value3), \log(value3) - \log(value2), \dots)$). A comparison of
131 the crowd forecasts against the default baseline shown in the application is displayed in Figure S25 in the
132 Appendix. Our interface also allowed participants to view past observations based on the hub data, as well as
133 their forecasts, on a logarithmic scale and presented additional contextual COVID-19 data sourced from (39).
134 These data included, for example, notifications of both test positive COVID-19 cases and COVID-19 linked
135 deaths and the number of COVID-19 tests conducted over time. From November 26 2020 on we displayed
136 weekly small reports with a visualisation of past forecasts and scores on our website, epiforecasts.io.

137 Forecasts were stored in a Google Sheet and downloaded, cleaned and processed every week for submission
138 to the Forecast Hub. If a forecaster had submitted multiple predictions for a single target, only the latest
139 submission was kept. Information on the chosen distribution as well as the parameters for median and width
140 were used to obtain the required set of 23 quantiles from that distribution. Forecasts from all forecasters were
141 then aggregated using an unweighted quantile-wise mean (i.e., the α -quantile of the ensemble is given by the
142 mean of all submitted α -quantiles). To avoid issues with users trying out the app and submitting a random
143 forecast, we required that a forecaster needed to make a forecast for at least two targets for a given forecast
144 in order to be included in the crowd forecast ensemble. On a few occasions ~~, we deleted forecasts that were~~
145 ~~clearly the result of a user or software error (such as for example forecasts that were zero everywhere).~~

146 individual forecasts were assessed as clearly erroneous by visual inspection and subsequently removed before
147 aggregation and were excluded from the submission as well as the analysis.

148 Participants were recruited mostly within the Centre of Mathematical Modeling of Infectious Diseases at the

¹⁴⁹ London School of Hygiene & Tropical Medicine, but participants were also invited personally or via social
¹⁵⁰ media to submit predictions. Depending on whether they had a background in either statistics, forecasting or
¹⁵¹ epidemiology, participants were asked to self-identify as ‘experts’ or ‘non-experts’.

¹⁵² The study was approved by the Observational / Interventions Research Ethics Committee at the London
¹⁵³ School of Hygiene & Tropical Medicine, LSHTM Ethics Reference: 22290.

¹⁵⁴ 4.3 Model-based forecasts

¹⁵⁵ We used two Bayesian semi-mechanistic models from the EpiNow2 R package (version 1.3.3) as our model-
¹⁵⁶ based forecasts (40). The first of these models, here called “renewal model”, used the renewal equation (41)
¹⁵⁷ to predict reported cases and deaths (see details in the SI). It estimated the effective reproduction number
¹⁵⁸ R_t (the average number of people each person infected at time t is expected to infect in turn) and modelled
¹⁵⁹ future infections as a weighted sum of past infection multiplied by R_t . R_t was assumed to stay constant
¹⁶⁰ beyond the forecast date, roughly corresponding to continuing the latest exponential trend in infections. On
¹⁶¹ the 9th of November we altered the date when R_t was assumed to be constant from two weeks prior to the
¹⁶² date of the forecast to the forecast date, which we found to yield a more stable R_t estimate. Reported case
¹⁶³ and death notifications were obtained by convolving predicted infections over data-based delay distributions
¹⁶⁴ (40,42–44) to model the time between infection and report date. The renewal model was used to predict cases
¹⁶⁵ as well as deaths with forecasts being generated for each target separately. Death forecasts from the renewal
¹⁶⁶ model were therefore not informed by past cases. One submission of the renewal model on December 28th
¹⁶⁷ 2020 was delayed and therefore not included in the official Forecast hub ensemble.

¹⁶⁸ The second model (“convolution model”, see details in SI) was only used to forecast deaths and was added
¹⁶⁹ later, starting December 7th 2020 (with the first forecast from December 7th suffering from a software bug
¹⁷⁰ and therefore disregarded in all further analyses). The convolution model was submitted, but never included
¹⁷¹ in the official Forecast hub ensemble due to concerns that it could be too similar to the renewal model. The
¹⁷² convolution model predicted deaths as a fraction of infected people who would die with some delay, by using
¹⁷³ a convolution of reported cases with a distribution that described the delay from case report to death and a
¹⁷⁴ scaling factor (the case-fatality ratio). Both the renewal and the convolution model used daily observations

175 and assumed a negative binomial observation model with a multiplicative day-of-the-week effect (40).

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

180 4.4 Analysis

181 For the main analysis we focused mostly on two week ahead forecasts, as COVID-19 forecasts, especially
182 for cases, were in the past found to have poor predictive performance beyond this horizon (15). Forecasts
183 for cases were scored using the full period from October 2020 until March 2021. To ensure comparability
184 between models, all death forecasts were scored using only the period from December 14th on, where all
185 models including the convolution model were available. To ensure robustness of our results we conducted a
186 sensitivity analysis where all forecasts (including cases) were scored only over the later period for which all
187 forecasts were available (see Section A.11 in the SI). Results remained broadly unchanged.

188 Forecasts were analysed using the following scoring metrics: The weighted interval score (WIS) (47), the
189 absolute error, relative bias, and empirical coverage of the 50% and 90% prediction intervals. The WIS
190 is a proper scoring rule (48), meaning that in expectation the score is optimised by reporting a predictive
191 distribution that is identical to the true data-generating distribution. Forecasters are therefore incentivised
192 to report their true belief about the future. The WIS can be understood as a generalisation of the absolute
193 error to quantile-based forecasts (also meaning that smaller values are better) and can be decomposed into
194 three separate penalties: forecast spread (i.e. uncertainty of forecasts), over-prediction and under-prediction.

195 While the over- and under-prediction components of the WIS capture the amount of over-prediction and
196 under-prediction in absolute terms, we also look at a relative tendency to make biased forecasts. The bias
197 metric (9) we use captures how much probability mass of the forecast was above or below the true value
198 (mapped to values between -1 and 1) and therefore represents a general tendency to over- or under-predict in
199 relative terms. A value of -1 implies that all quantiles of the predictive distribution are below the observed
200 value and a value of 1 that all quantiles are above the observed value. Empirical coverage is the percentage

201 of observed values that fall inside a given prediction interval (e.g. how many observed values fall inside all
202 50% prediction intervals). Scoring metrics are explained in more detail in Table S1 in the SI. All scores were
203 calculated using the **scoringutils** R package (49).

204 At all stages of the evaluation our forecasts were compared to the median ensemble of all *other* models
205 submitted to the German and Polish Forecast Hub (“Hub ensemble”). This “Hub ensemble” was retrospectively
206 computed and excludes all our models.~~—It~~, leaving on average five ensemble member models (see details in
207 [Section 5.4, as well as in Table S10 and Figure S24 in the SI](#)). What we call “Hub ensemble” in this article
208 therefore differs from the “[official Hub ensemble](#)” (here called “hub-ensemble-realised”) which included crowd
209 forecasts as well as renewal model forecasts. To enhance interpretability of scores we mainly report WIS
210 relative to the Hub ensemble in the main text, i.e. we divided the average scores for a given model by
211 the average score achieved by the Hub ensemble on the same set of forecasts (with values >1 implying
212 worse and values <1 implying better performance than the Hub ensemble). In addition to comparing our
213 forecasts against the hub ensemble excluding our models, we also assessed the impact of our forecasts on the
214 performance of the forecasting hub by recalculating separate versions of the Hub ensemble with only some
215 (or all) of our forecasts included. Versions that included either all of our models (“hub-ensemble-with-all”) or
216 only one of them (“hub-ensemble-with-X”) were computed retrospectively.

217 5 Results

218 5.1 Crowd forecast participation

219 A total number of 32 participants submitted forecasts, 17 of those self-identified as ‘expert’ in either forecasting
220 or epidemiology. The median number of forecasters for any given forecast target was 6, the minimum 2 and
221 the maximum 10. The mean number of submissions from an individual forecaster was 4.7 but the median
222 number was only one - most participants dropped out after their first submission. Only two participants
223 submitted a forecast every single week, both of whom are authors on this study.

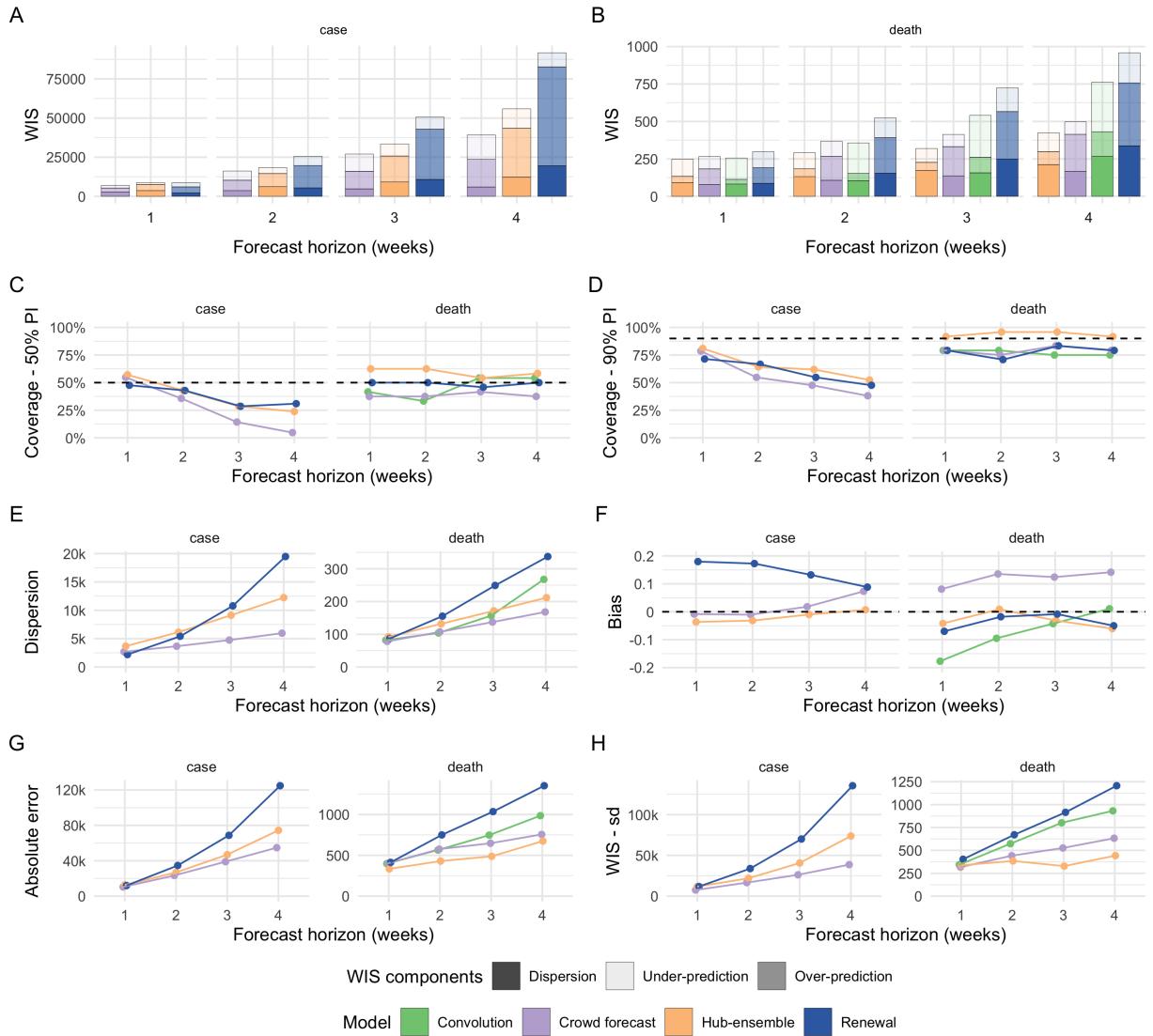


Figure 1: Visualisation of aggregate performance metrics **aeross**—for forecasts one to four weeks into the future(**marked as 1–4 on the x-axis**). A, B: mean weighted interval score (WIS, lower indicates better performance) across horizons. WIS is decomposed into its components dispersion, over-prediction 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

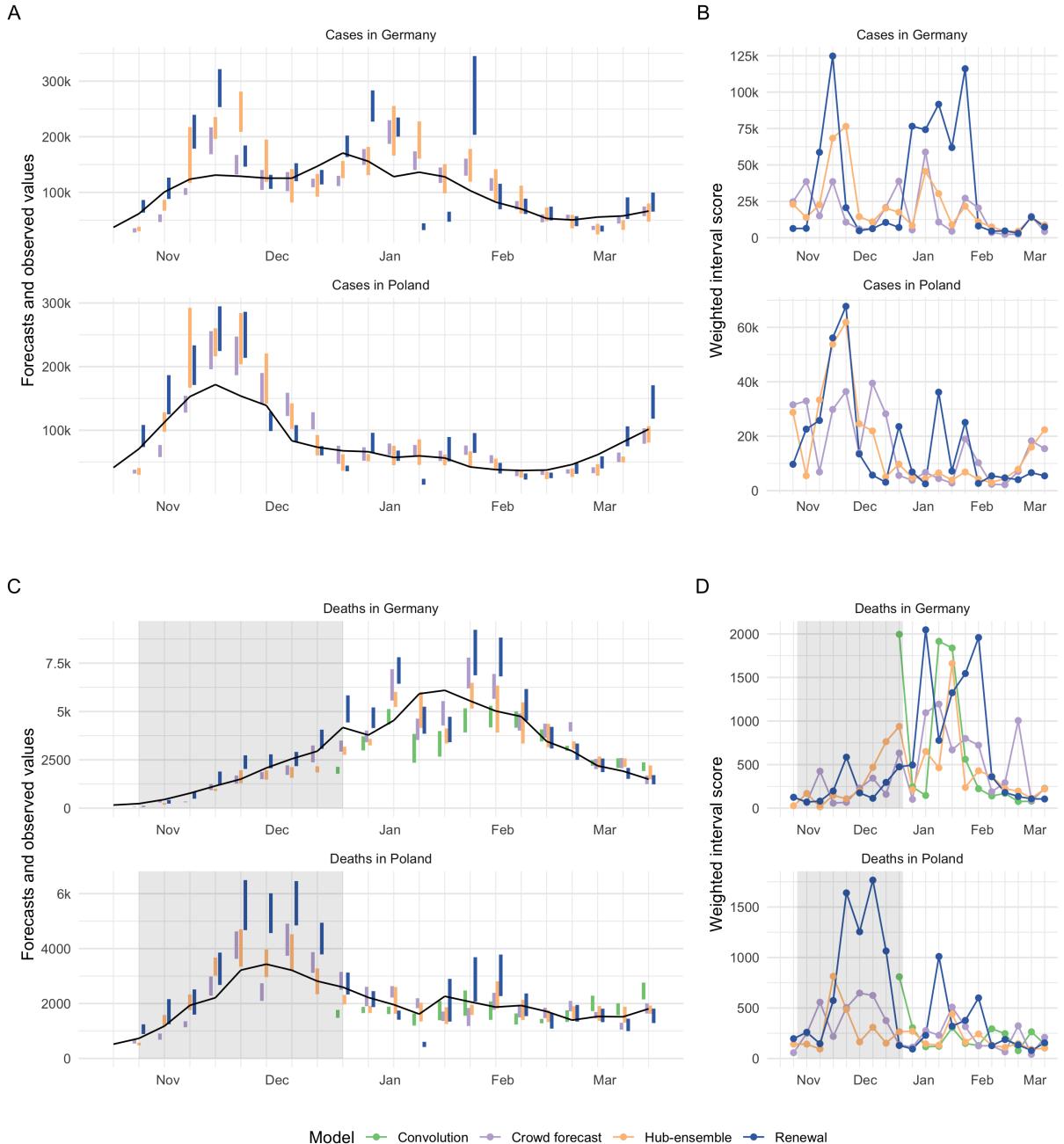


Figure 2: A, C: Visualisation of 50% prediction intervals of two week ahead forecasts against the reported values. Forecasts that were not scored (because there was no complete set of death forecasts available) are greyed out. B, D: Visualisation of corresponding WIS.

224 5.2 Case Forecasts

225 For cases, crowd forecasts had a lower mean weighted interval score (WIS, lower values indicate better
226 performance) than both the renewal model and the Hub ensemble across all forecast horizons (Figure 1A) and
227 locations (Figure S5A). For two week ahead forecasts, mean WIS relative to the Hub ensemble (= 1) was 0.89
228 for crowd forecasts and 1.40 for the renewal model (Table S2). Across all forecasting approaches, locations
229 and forecast horizons, the distribution of WIS values was very right-skewed, and average performance was
230 heavily influenced by outliers (Figure 3). Overall, low variance in forecast performance was closely linked with
231 good mean performance (Figures 1H and and 1A), suggesting that the ability to avoid large errors was an
232 important factor in determining overall performance. The impact of outlier values was especially pronounced
233 for the renewal model, which had more outliers (Figure 3), as well as the highest standard deviation of WIS
234 values (standard deviation of the WIS relative to the WIS sd of the Hub ensemble was 1.54 at the two weeks
235 ahead horizon), while the ensemble of crowd forecasts (rel. WIS sd 0.76) and the Hub ensemble (= 1) showed
236 more stable performance.

237 To varying degrees, all forecasts exhibited trend-following behaviour and were rarely able to predict a change
238 in trend before it had happened. For example, all forecasts failed to predict the change in trend from increase
239 to decrease that happened in November in Germany and severely overshot reported cases (Figure 2A). This
240 was most striking for the renewal model, which extrapolated unconstrained exponential growth based on the
241 recent past of observations. The Hub ensemble and the crowd forecast, which had both been under-predicting
242 throughout October, also failed to predict the change in trend after cases peaked, but less severely so. Human
243 forecasters, possibly aware of the semi-lockdown announced on November 2nd 2020 (50) and the change in the
244 testing regime (with stricter test criteria) on November 11th 2020 (36), were fastest to adapt to the new trend,
245 and the Hub ensemble slowest. In December, cases rose again in Germany, with all models under-predicting
246 this growth to varying extents. As in October, the renewal model captured the phase of exponential growth
247 in cases slightly better than other approaches, but again overshot when reported case numbers fell over
248 Christmas. The large variance in predictions in January in Germany (severe under-prediction followed by
249 severe over-prediction) may in part be caused by the fact that the renewal model operated on daily data and
250 therefore was susceptible to fluctuations in daily reporting around Christmas that would not have influenced

251 on weekly reporting. Similar trends in performance were evident in Poland, with the crowd forecast quickest
252 at adapting to the change in trend in November. In general, there were fewer large outlier forecasts in Poland
253 and in particular the renewal model performed more in line with other forecasts there.

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

263 In spite of good performance in terms of the absolute error (Figure 1G), the narrow forecast intervals led to
264 forecasts which were severely overconfident (covering only 36% and 55% of all observations with the 50% and
265 90% prediction intervals of all forecasts made at a two week forecast horizon, and only 5% and 38% four
266 weeks ahead) (Figure 1C,D and Tables S2 and S3). Despite worse performance in terms of absolute error
267 (Figure 1G), the renewal model achieved better calibration (comparable to the Hub ensemble), as uncertainty
268 increased rapidly across forecast horizons. The crowd forecasts, on the other hand, showed a smaller bias
269 than the renewal model, but were overconfident.

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

276 Generally, trends in overall performance were broadly similar across locations (Figures S4 and S5). Due

277 to the differing population sizes and numbers of notifications in Germany and Poland absolute scores were
278 difficult to compare directly. However, relative to the Hub ensemble, the crowd forecasts performed noticeably
279 better in Germany than in Poland and the renewal model better in Poland than in Germany ([Figure S5A](#) and
280 [S5G](#)[Figures S5A, S5G, S2, S3](#)).

281 5.3 Death Forecasts

282 For deaths, the Hub ensemble outperformed the crowd forecasts as well as our model-based approaches across
283 all forecast horizons and locations (Figure 1B, Figure S4B). Relative WIS values for the models two weeks
284 ahead were 1.22 (convolution model), 1.26 (crowd forecast), 1 (Hub ensemble) and 1.79 (renewal model). The
285 crowd forecasts performed better than the renewal model across all forecast horizons and locations (Figure
286 1B, Figure S4B), and also better than the convolution model three and four weeks ahead. Poor performance
287 of the renewal model, especially at longer horizons, indicates that an approach that does not know about
288 past cases, but instead estimates and projects a separate R_t trace from deaths, does not use the available
289 information efficiently. The convolution model was able to outperform both the renewal model and the crowd
290 forecasts at shorter forecast horizons (where the delay between cases and deaths means that future deaths are
291 largely informed by present cases), but saw performance deteriorate at three and four weeks ahead (where
292 case predictions from the renewal model were increasingly used to inform death predictions) (Figure 1B,
293 Table S3).

294 As past cases and hospitalisations can be used as predictors, predicting a change in trend may be easier for
295 deaths than for cases. Even though all forecasts generally struggled with this, there were some instances
296 where changing trends were well captured or even anticipated. In Poland, for example, the Hub ensemble was
297 able to capture or even anticipate the peak in deaths in December quite well (whereas the renewal model and
298 crowd forecast did not). The renewal model, which mostly exhibited trend-following behaviour, correctly
299 predicted another increase in weekly deaths in mid-January (potentially based on changes in daily deaths,
300 as the renewal model did not know about past cases). In Germany in early January, all models predicted
301 a decrease in deaths two to three weeks before it actually happened. Predictions from the renewal model
302 at that time were likely strongly influenced by an unexpected drop in reported deaths in December. The

303 other forecasting approaches and in particular, the convolution model may have been affected by potentially
304 under-reported case numbers around Christmas. When the decrease that all models had predicted to happen
305 in early January failed to materialise, the renewal model and the crowd forecast noticeably over-corrected
306 and over-predicted deaths in the following weeks, while the Hub ensemble, and to a slightly lesser degree, the
307 convolution model were able to capture the downturn well when it finally happened at the end of January.

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

318 For deaths, the ensemble of crowd forecasts had a consistent tendency to over-predict 1F. The convo-
319 lution model had a strong tendency to under-predict, ~~which steadily decreased with the magnitude of~~
320 under-prediction steadily decreasing for longer forecast horizons. The renewal model (which over-predicted
321 for cases) and the Hub ensemble slightly tended towards under-prediction. For deaths, absolute over- and
322 under-prediction penalties were more in line with a general relative tendency to over- or under-predict than
323 for cases (Figure 1A, 1B and Tables S2, S3).

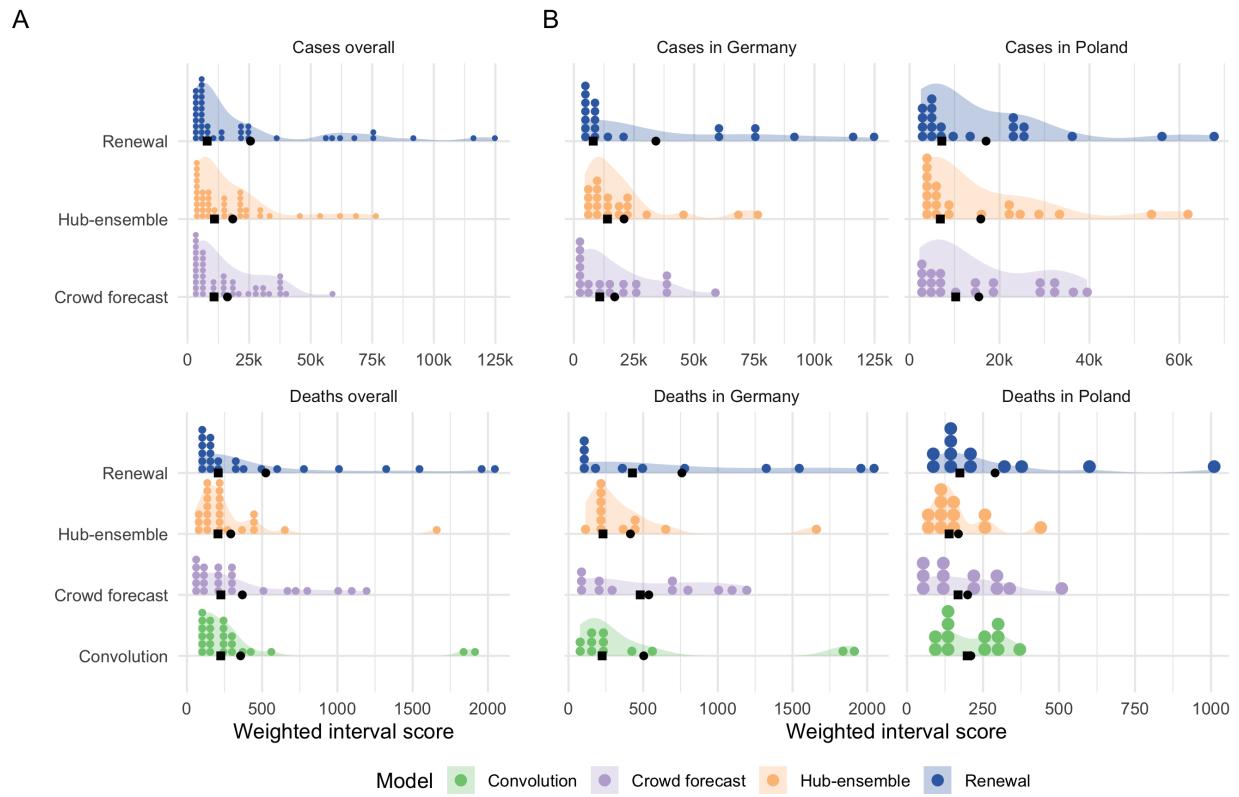


Figure 3: A: Distribution of weighted interval scores for two week ahead forecasts of the different models and forecast targets. Points denote single forecasts scores, while the shaded area shows an estimated probability density. B: Distribution of WIS separate by country. Black squares indicate median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

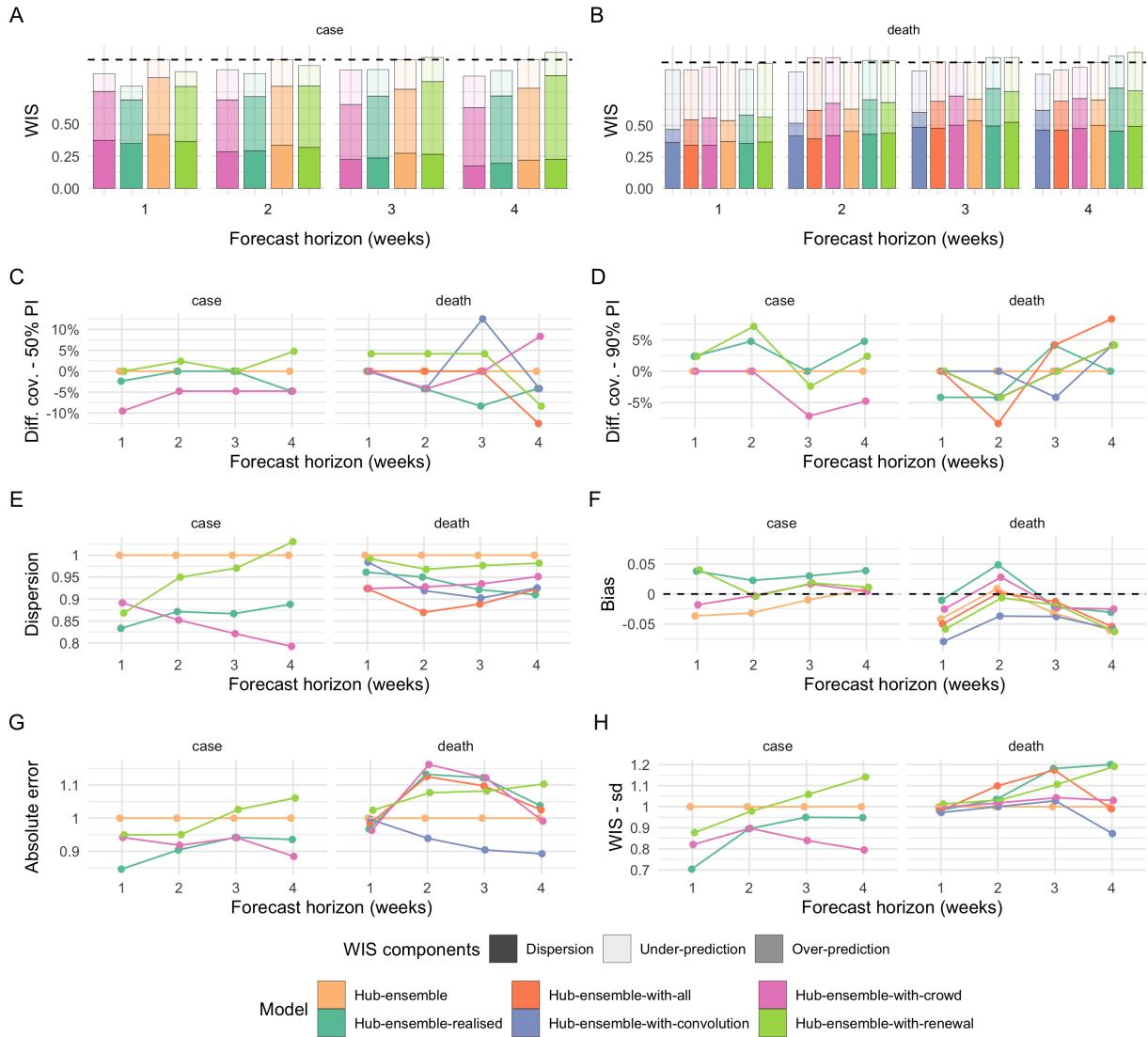


Figure 4: Visualisation of relative aggregate performance metrics across forecast horizons for the different versions of the Hub median ensemble. “Hub-ensemble” extitexcludes all our models, Hub-ensemble-all extitincludes all of our models, “Hub-ensemble-realHub-ensemble-realised” is the real-actual hub-ensemble with-observed in reality, which includes the renewal model and the crowd forecastsincluded. Values (except for Bias) are computed as differences to but not the Hub ensemble excluding our contributionsconvolution model. For Coverage, this is an absolute difference, for other metrics this is a percentage difference. A, B: mean weighted interval score (WIS) across horizons. C: Absolute error of relative to the median forecastHub ensemble (lower values indicate better performance). D: Standard deviation Empirical coverage of the WIS50% and 90% prediction intervals minus empirical coverage observed for the Hub ensemble. E: Dispersion (higher relative to the dispersion of the Hub ensemble. Higher values mean greater spread dispersion of the forecast.) and imply ceteris paribus a worse score. F: Bias, i.e., general

³²⁴ 5.4 Contribution to the Forecast Hub

³²⁵ Of our three models, only the renewal model and the crowd forecast were included in the official Forecast Hub
³²⁶ median ensemble (“hub-ensemble-realised”), while the convolution model was never included as it was deemed
³²⁷ too similar to the existing renewal model. In the official Hub ensemble, there were on average 7.1 models
³²⁸ included (including our own), with a median of 7, a minimum of 4 (on 28 December 2020 over the Christmas
³²⁹ period) and a maximum of 10. Versions that included either all of our models (“hub-ensemble-with-all”) or
³³⁰ only one of them (“hub-ensemble-with-X”) were computed retrospectively. An overview of all models and
³³¹ ensemble versions is shown in Table S10 in the SI.

³³² For cases, our contributions (compared to the Hub ensemble without our contributions) consistently improved
³³³ performance across all forecasting horizons (rel. WIS 0.9 two weeks ahead, Table S4). Contributions from
³³⁴ the crowd forecasts alone also improved performance of the Hub ensemble across all forecast horizons, while
³³⁵ contributions from the renewal model had a negative effect for longer horizons (rel. WIS 1.02 three weeks
³³⁶ ahead, 1.06 four weeks ahead). The realised ensemble including both models performed better or equal
³³⁷ compared to all versions with only one model included for up to three weeks ahead, suggesting synergistic
³³⁸ effects. Only for predictions four weeks ahead would removing the renewal model have improved performance
³³⁹ (Table S5). The realised ensemble performed comparably to the crowd forecasts for predictions one to two
³⁴⁰ weeks ahead, and worse for greater forecast horizons.

³⁴¹ For deaths, contributions from the renewal model and crowd forecast together improved performance only for
³⁴² one week ahead predictions and showed an increasingly negative impact on performance for longer horizons
³⁴³ (rel. WIS of the hub-ensemble-realised-Hub-ensemble-realised 1.01 two weeks ahead, 1.05 four weeks ahead,
³⁴⁴ Tables S4 and S5). Individual contributions from both the renewal model and the crowd forecast were largely
³⁴⁵ negative, while a version of the Hub ensemble with only the convolution model included would have performed
³⁴⁶ consistently better across all forecast horizons (with the positive impact increasing for longer horizons). This
³⁴⁷ is especially interesting as the convolution model performed consistently worse than the pre-existing Hub
³⁴⁸ ensemble (Figure 1) and especially worse for longer horizons.

³⁴⁹ We also considered the impact of our contributions on a version of the Hub ensemble constructed by taking

350 the quantile-wise mean, rather than the median. General trends were similar, with the notable exception of
351 the convolution model, which had a consistently positive impact on the median ensemble, but a mixed and
352 mostly slightly negative impact on the mean ensemble (Figures 224B and S21B). This may happen if a model
353 is more correct directionally relative to the pre-existing ensemble, but overshoots in absolute terms, thereby
354 moving the ensemble too far. For both the mean and the median ensemble, changes in performance from
355 adding or removing models were of a similar order of magnitude, suggesting that at least in this instance,
356 with a relatively small ensemble size, the median ensemble was not necessarily more ‘robust’ to changes than
357 the mean ensemble. However, the ensemble version with all our forecasts included (“hub-ensemble-with-all”)
358 tended to perform relatively better for the median ensemble than the mean ensemble, suggesting that adding
359 more models may be more beneficial or ‘safer’ for the median than for the mean ensemble as directional
360 errors can more easily cancel out than errors in absolute terms.

361 6 Discussion

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

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

375 (11), but extends it in several ways. While Farrow et al. only asked for point predictions and constructed a
376 predictive distribution from these, we asked participants to provide a full predictive distribution, allowing
377 us to compare human forecasts and models without any further assumptions, as well as to analyse how
378 humans quantified their uncertainty. In addition, we compared crowd forecasts to two semi-mechanistic
379 models informed by basic epidemiological knowledge of COVID-19, allowing us to assess not only relative
380 performance but also to analyse qualitative differences between human judgement and model-based insight.
381 In terms of interpretability of the results, exact knowledge of our two models, as well as focus on a limited
382 set of targets and locations was a major advantage of our study compared to larger studies conducted by the
383 Forecast Hubs (12–15,17).

384 The strong performance of crowd forecasts in our study is in line with results from Farrow et al. who also report
385 strong performance of human predictions in past Flu challenges despite difficulties to recruit a large number
386 of participants. The advantage of crowd forecasts we observed over our semi-mechanistic models is likely in
387 part explained by the fact that we compared an ensemble of crowd forecasts with single models. However, this
388 probably explains only part of the difference, and performance relative to the Hub ensemble strongly suggests
389 that human insight is valuable when forecasting highly volatile and potentially hard-to-predict quantities such
390 as case numbers. One potential explanation is that humans can have access to data that is not available to or
391 hard to integrate into model-based forecasts. Relatively good performance of our semi-mechanistic models
392 short-term, but not longer-term, suggests that model-based forecasts are helpful to extrapolate from current
393 conditions, but require some form of human intervention or additional assumptions to inform forecasts when
394 conditions change over time. This human intervention may be particularly important when dealing with
395 artefacts in reporting and data anomalies (and especially when using daily, rather than weekly data). The
396 large variance in predictions in January in Germany for example (severe under-prediction followed by severe
397 over-prediction, see Figure 2A), may in part be caused by the fact that the renewal model operated on daily
398 data and therefore was susceptible to fluctuations in daily reporting ~~around Christmas that would not have~~
399 ~~influenced which have less of an influence~~ on weekly reporting.

400 Our results suggest that human intervention may be less beneficial when forecasting deaths (especially at
401 shorter horizons, when deaths are largely dependent on already observed cases), which benefits from the ability

402 to model the delays and exact epidemiological relationships between different leading and lagged indicators.

403 Relatively good performance of the convolution model, especially compared to the poor performance of the

404 renewal model on deaths (which used only deaths to estimate and predict the effective reproduction number)

405 underlines the importance of including leading indicators such as cases as a predictor for deaths.

406 Given the low number of participants in our study, it is difficult to generalise conclusions about crowd

407 predictions to other settings. Using R shiny as a platform for the web application arguably created some

408 limits to user experience and performance, influencing the number of participants and potentially creating

409 a self-selection effect. Motivating forecasters to contribute regularly proved challenging, especially given

410 that the majority of our participants were from the UK and may not have been familiar with all relevant

411 details of the situation in Germany and Poland. On the other hand, R shiny facilitated quick development

412 and allowed us to provide our crowd forecasting tooling as an open source R package, meaning that it is

413 available for others to use, for example in settings like early-stage outbreaks where model-based forecasts

414 are not available. In light of the relatively small number of Hub ensemble models, performance of the Hub

415 ensemble is also difficult to generalise. More research is needed to replicate these findings and investigate

416 how crowd forecasts compare against the types of models and model ensembles policy makers use to inform

417 their decisions.

418 Our work suggests that crowd forecasts and model-based forecasts could have different strengths and may

419 be able to complement each other. When choosing a suitable approach for a given task it is important to

420 take into account how the output will be used. In this work we focused on forecasts (which aim to predict

421 future data points whilst accounting for all factors that might influence them), whereas policy makers might

422 be more interested in projections (which show what would happen in the absence of any events that could

423 change the trend) or scenario modelling. Forecasts may not be a suitable basis for informing policy decisions,

424 if forecasters already have factored in the expectation of a future intervention. Model-based approaches can

425 be either forecasts or projections depending on the assumptions, whereas eliciting projections that are not

426 influenced by implicit assumptions about the future from humans may be harder.

427 Further work should explore the effects of humans refining their mathematical models or changing model

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

440 **Acknowledgements**

441 NIB received funding from the 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
442 work of JB was supported by the Helmholtz Foundation (<https://www.helmholtz.de/>) via the SIMCARD
443 Information and Data Science Pilot Project. This research was partly funded by the National Institute
444 for Health Research (NIHR, <https://www.nihr.ac.uk/>) (16/137/109 & 16/136/46) using UK aid from the
445 UK Government to support global health research. The views expressed in this publication are those
446 of the author(s) and not necessarily those of the NIHR or the UK Department of Health and Social
447 Care. BJQ is supported in part by a grant from the Bill and Melinda Gates Foundation (OPP1139859,
448 <https://www.gatesfoundation.org/>). EvL acknowledges funding by the National Institute for Health Research
449 (NIHR) Health Protection Research Unit (HPRU) in Modelling and Health Economics (grant number
450 NIHR200908, <https://www.nihr.ac.uk/>) and the European Union's Horizon 2020 research and innovation
451 programme - project EpiPose (101003688, <https://ec.europa.eu/programmes/horizon2020/>). AC acknowledges
452 funding by the NIHR, the Sergei Brin foundation, USAID (<https://www.usaid.gov/>), and the Academy
453 of Medical Sciences (<https://acmedsci.ac.uk/>). SF's work was supported by the Wellcome Trust (grant:
454 210758/Z/18/Z, <https://wellcome.org/>), and the NIHR (NIHR200908, <https://www.nihr.ac.uk/>).

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

457 We would also like to acknowledge (in a randomised order) the members of Centre for the Mathematical
458 Modelling of Infectious Diseases COVID-19 Working Group at the the London School of Hygiene & Tropical
459 Medicine: Oliver Brady, Katharine Sherratt, Kaja Abbas, Kerry LM Wong, Charlie Diamond, Katherine
460 E. Atkins, Rein M G J Houben, Jiayao Lei, Rachel Lowe, David Simons, Sophie R Meakin, Nicholas G.
461 Davies, Timothy W Russell, Kevin van Zandvoort, Quentin J Leclerc, Kathleen O'Reilly, Stéphane Hué,
462 Alicia Rosello, Emilie Finch, C Julian Villabona-Arenas, Thibaut Jombart, W John Edmunds, Yalda Jafari,
463 Jack Williams, Alicia Showering, Damien C Tully, Jon C Emery, Carl A B Pearson, David Hodgson, Frank
464 G Sandmann, Petra Klepac, Adam J Kucharski, Graham Medley, Yang Liu, Simon R Procter, Emily S
465 Nightingale, William Waites, Rosanna C Barnard, Joel Hellewell, Yung-Wai Desmond Chan, Fiona Yueqian

⁴⁶⁶ Sun, Hamish P Gibbs, Rosalind M Eggo, Lloyd A C Chapman, Stefan Flasche, James W Rudge, Akira
⁴⁶⁷ Endo, Naomi R Waterlow, Paul Mee, James D Munday, Ciara V McCarthy, Mihaly Koltai, Amy Gimma,
⁴⁶⁸ Christopher I Jarvis, Megan Auzenberg, Matthew Quaife, Fabienne Krauer, Samuel Clifford, Georgia R
⁴⁶⁹ Gore-Langton, Arminder K Deol, Kiesha Prem, Gwenan M Knight, Rachael Pung, Anna M Foss.

⁴⁷⁰ **A Supplementary information**

⁴⁷¹ **A.1 Scoring metrics used**

Table S1: Overview of the scoring metrics used.

Metric	Explanation
WIS (Weighted interval score)	<p>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</p> $IS_\alpha(F, y) = (u - l) + \frac{2}{\alpha} \cdot (l - y) \cdot 1(y \leq l) + \frac{2}{\alpha} \cdot (y - u) \cdot 1(y \geq u),$ <p>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,</p> $WIS = \frac{1}{K + 0.5} \cdot \left(w_0 \cdot y - m + \sum_{k=1}^K w_k \cdot IS_\alpha(F, y) \right),$ <p>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.</p>

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

Metric	Explanation
Interval coverage	<p>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.</p>
Bias	<p>(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</p> $B(F, y) = 1 - 2 \cdot (F(y)),$ <p>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.</p>

472 A.2 The crowdforecasting app

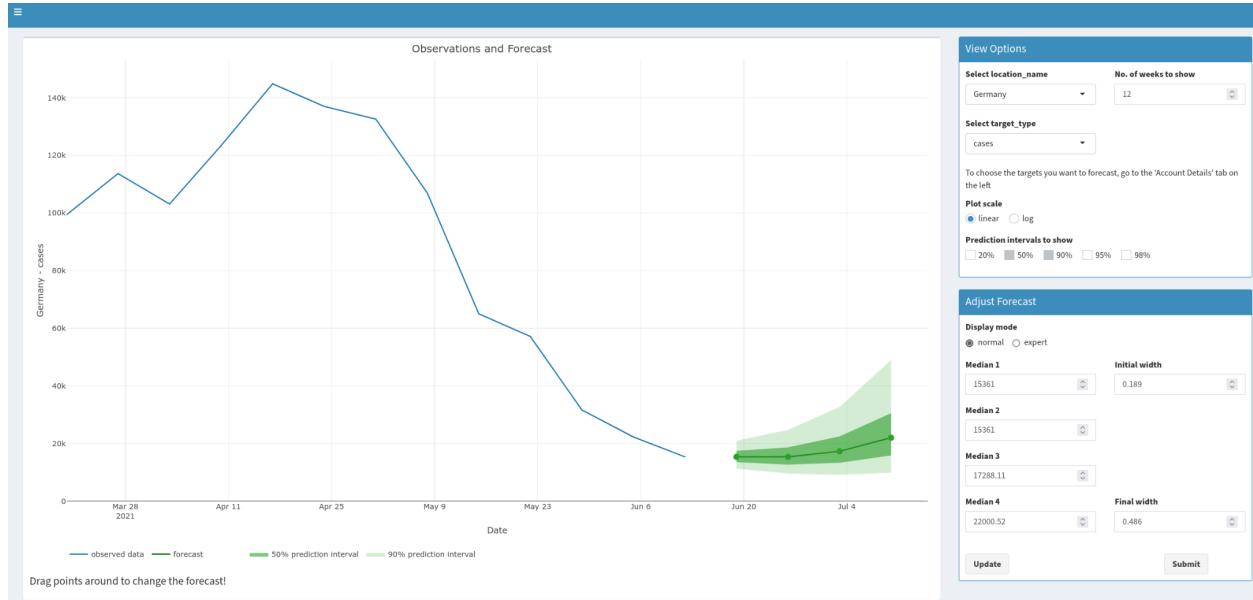


Figure S1: Screenshot of the crowdforecasting app used to elicit predictions (made in June 2021).

473 **A.3 Further details on the semi-mechanistic forecasting models**

474 **A.3.1 Renewal equation model**

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

$$I_t = I_0 \exp(rt) \quad (1)$$

$$I_0 \sim \mathcal{LN}(\log I_{obs}, 0.2) \quad (2)$$

$$r \sim \mathcal{LN}(r_{obs}, 0.2) \quad (3)$$

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

$$\log R_t = \log R_{t-1} + \text{GP}_t \quad (4)$$

$$I_t = R_t \sum_{\tau=1}^{15} w(\tau | \mu_w, \sigma_w) I_{t-\tau} \quad (5)$$

$$O_t = \sum_{\tau=0}^{15} \xi_O(\tau | \mu_{\xi_O}, \sigma_{\xi_O}) I_{t-\tau} \quad (6)$$

$$D_t = \alpha \sum_{\tau=0}^{15} \xi_D(\tau | \mu_{\xi_D}, \sigma_{\xi_D}) O_{t-\tau} \quad (7)$$

$$C_t \sim \text{NB}(\omega_{(t \mod 7)} D_t, \phi) \quad (8)$$

Where,

$$w \sim \mathcal{G}(\mu_w, \sigma_w) \quad (9)$$

$$\xi_O \sim \mathcal{LN}(\mu_{\xi_O}, \sigma_{\xi_O}) \quad (10)$$

$$\xi_D \sim \mathcal{LN}(\mu_{\xi_D}, \sigma_{\xi_D}) \quad (11)$$

⁴⁸² This model used the following priors for cases,

$$R_0 \sim \mathcal{LN}(0.079, 0.18) \quad (12)$$

$$\mu_w \sim \mathcal{N}(3.6, 0.7) \quad (13)$$

$$\sigma_w \sim \mathcal{N}(3.1, 0.8) \quad (14)$$

$$\mu_{\xi_O} \sim \mathcal{N}(1.62, 0.064) \quad (15)$$

$$\sigma_{\xi_O} \sim \mathcal{N}(0.418, 0.069) \quad (16)$$

$$\mu_{\xi_D} \sim \mathcal{N}(0.614, 0.066) \quad (17)$$

$$\sigma_{\xi_D} \sim \mathcal{N}(1.51, 0.048) \quad (18)$$

$$\alpha \sim \mathcal{N}(0.25, 0.05) \quad (19)$$

$$\frac{\omega}{7} \sim \text{Dirichlet}(1, 1, 1, 1, 1, 1, 1) \quad (20)$$

$$\phi \sim \frac{1}{\sqrt{\mathcal{N}(0, 1)}} \quad (21)$$

⁴⁸³ and updated the reporting process as follows when forecasting deaths,

$$\mu_{\xi_D} \sim \mathcal{N}(2.29, 0.076) \quad (22)$$

$$\sigma_{\xi_D} \sim \mathcal{N}(0.76, 0.055) \quad (23)$$

$$\alpha \sim \mathcal{N}(0.005, 0.0025) \quad (24)$$

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

485 The prior for the generation time was sourced from (51) but refit using a log-normal incubation period with
486 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
487 a prior (52) for ξ_O . This resulted in a gamma-distributed generation time with mean 3.6 days (standard
488 deviation (SD) 0.7), and SD of 3.1 days (SD 0.8) for all estimates. We estimated the delay between symptom
489 onset and case report or death required to convolve latent infections to observations by fitting an integer
490 adjusted log-normal distribution to 10 subsampled bootstraps of a public linelist for cases in Germany from
491 April 2020 to June 2020 with each bootstrap using 1% or 1769 samples of the available data (45,53) and
492 combining the posteriors for the mean and standard deviation of the log-normal distribution (40,42,46,54).

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

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

$$I_t = (N - I_{t-1}^c) \left(1 - \exp \left(\frac{-I'_t}{N - I_T^c} \right) \right), \quad (25)$$

501 where $I_t^c = \sum_{s < t} I_s$ are cumulative infections by $t - 1$ and I'_t are the unadjusted infections defined above.

502 This adjustment is based on that implemented in the `epidemia` R package (56,57).

503 **A.3.1.1 Convolution model** The convolution model shares the same observation model as the renewal
504 model but rather than assuming that an observation is predicted by itself using the renewal equation instead
505 assumes that it is predicted entirely by another observation after some parametric delay. It can be defined
506 mathematically as follows,

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

507 with the following priors,

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

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

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

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

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

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

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

509 **A.3.2 Model fitting**

510 Both models were implemented using the `EpiNow2` R package (version 1.3.3) (40). Each forecast target was
511 fitted independently for each model using Markov-chain Monte Carlo (MCMC) in `stan` (46). A minimum of 4
512 chains were used with a warmup of 250 samples for the renewal equation-based model and 1000 samples for
513 the convolution model. 2000 samples total post warmup were used for the renewal equation model and 4000

514 samples for the convolution model. Different settings were chosen for each model to optimise compute time
515 contingent on convergence. Convergence was assessed using the R hat diagnostic (46). For the convolution
516 model forecast the case forecast from the renewal equation model was used in place of observed cases beyond
517 the forecast horizon using 1000 posterior samples. 12 weeks of data was used for both models though only 3
518 weeks of data were included in the likelihood for the convolution model.

⁵¹⁹ **A.4 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
2 wk ahead	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
	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
Renewal	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
4 wk ahead	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
	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

520 **A.5 Aggregate performance by location**

521 **A.5.1 Performance in Germany**

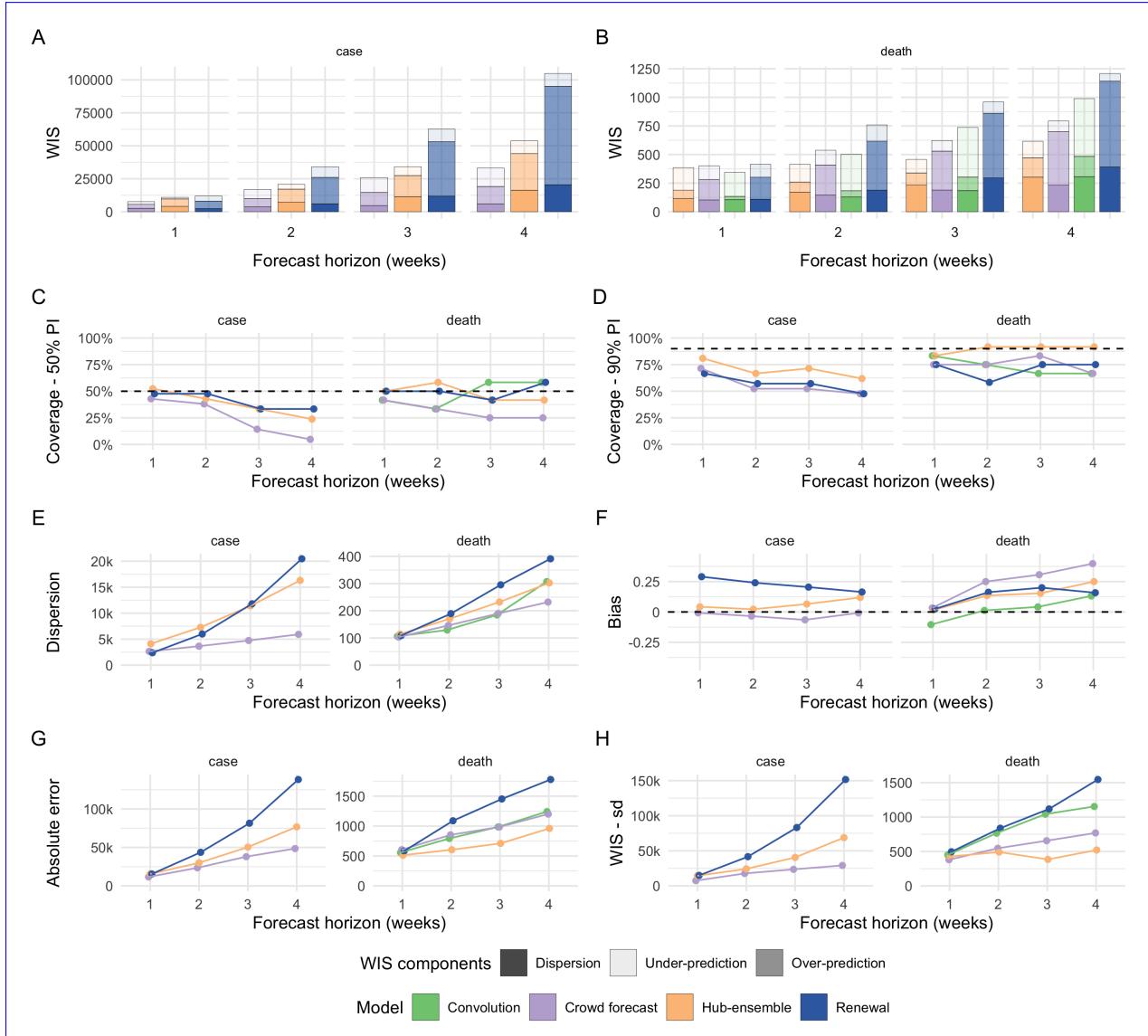


Figure S2: [Visualisation of aggregate performance metrics for forecasts one to four weeks into the future in Germany](#). A, B: mean weighted interval score (WIS, lower indicates better performance) across horizons. WIS is decomposed into its components dispersion, over-prediction and under-prediction. 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

522 A.5.2 Performance in Poland

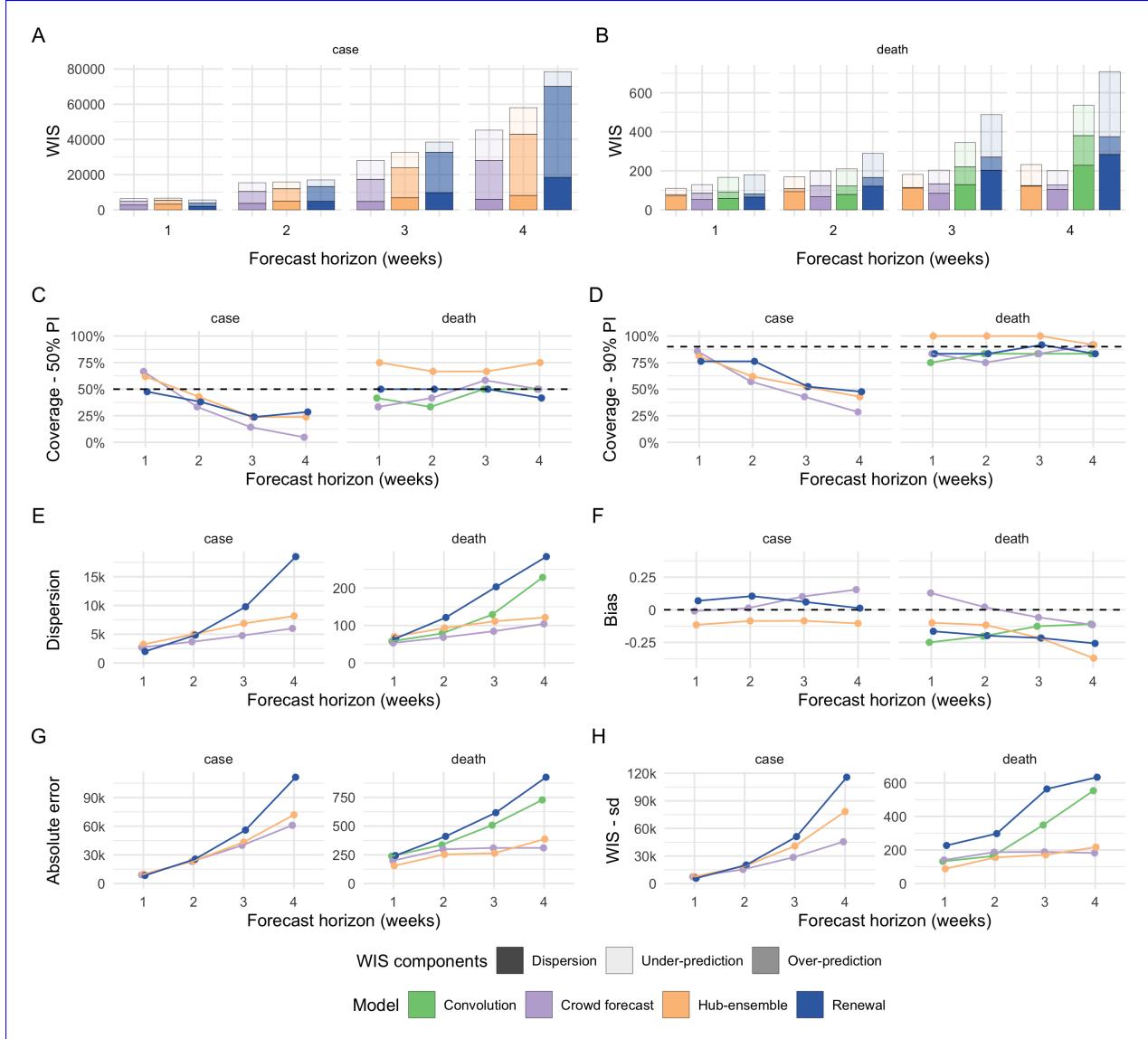


Figure S3: [Visualisation of aggregate performance metrics for forecasts one to four weeks into the future in Poland](#). A, B: mean weighted interval score (WIS, lower indicates better performance) across horizons. WIS is decomposed into its components dispersion, over-prediction and under-prediction. 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

523 A.5.3 Performance across locations in absolute terms

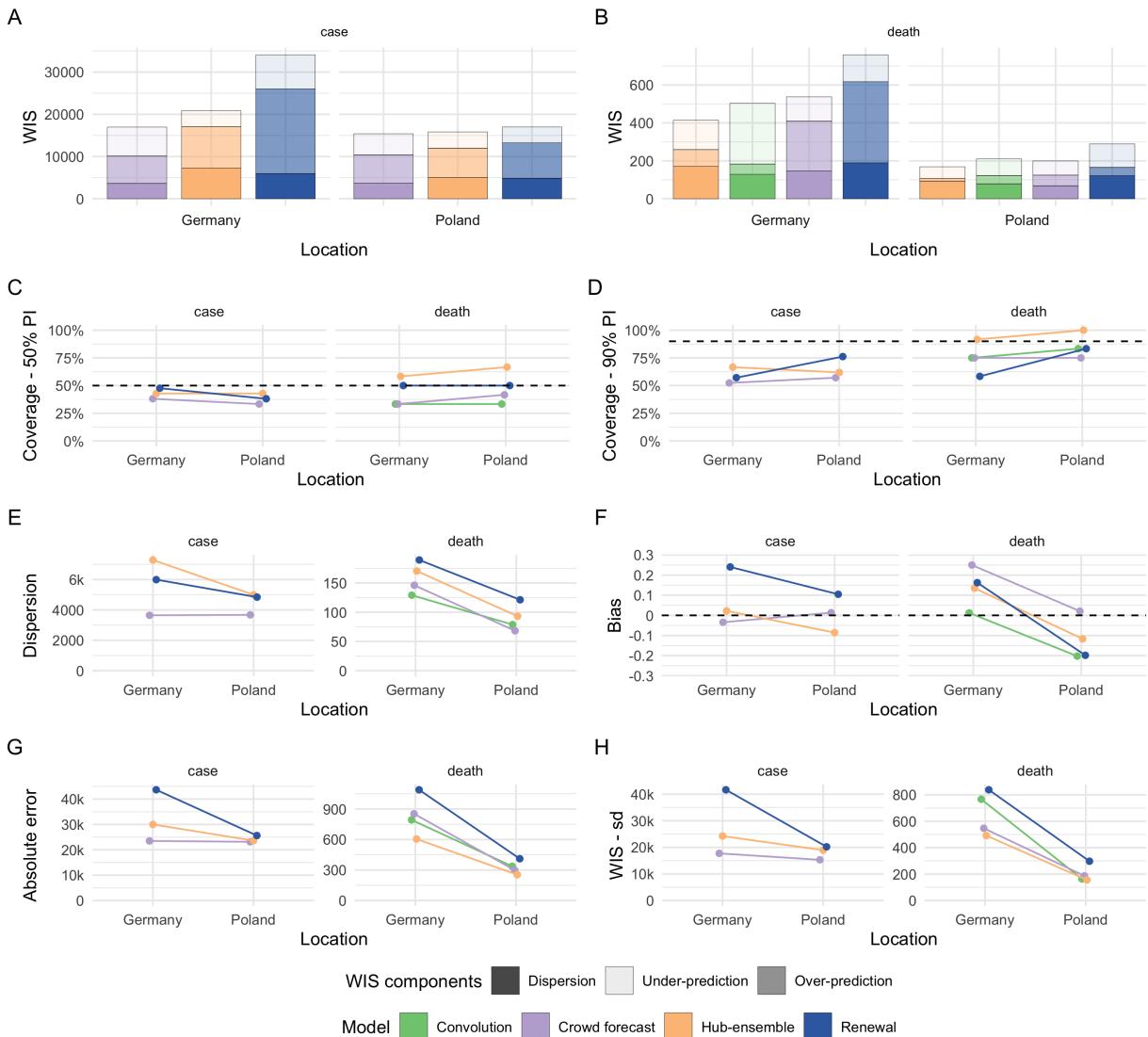


Figure S4: Visualisation of aggregate performance metrics across locations. A, B: mean weighted interval score (WIS, lower indicates better performance) across horizons. B: median WIS is decomposed into its components dispersion, over-prediction and under-prediction. C: Absolute error Empirical coverage of the median forecast 50% prediction intervals (50% coverage is perfect). D: Standard deviation Empirical coverage of the WIS 90% prediction intervals. E: Dispersion (higher same as in panel A, B). Higher values mean further spread out 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: Empirical coverage Absolute error of the 50% prediction intervals median forecast (lower is better). F: Empirical coverage H: Standard deviation of the 90% prediction intervals WIS values.

524 A.6 Performance across locations in relative terms

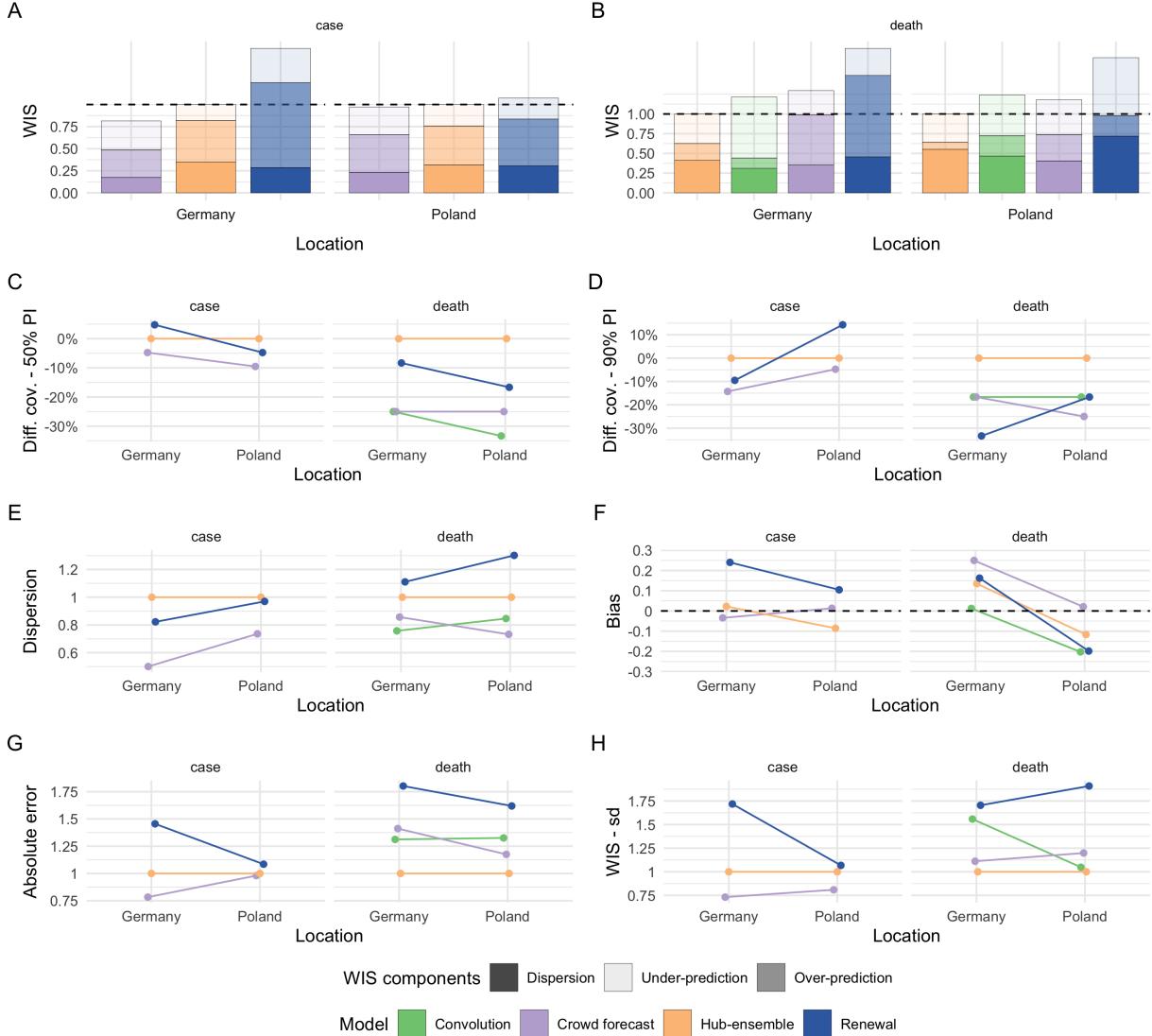


Figure S5: Visualisation of relative aggregate performance metrics across locations relative to the Hub ensemble (excluding our contributions). A, B: mean weighted interval score (WIS) across horizons locations (lower values indicate better performance). B: median WIS. C: Absolute error of the median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

525 A.7 Visualisation of daily reported cases and deaths

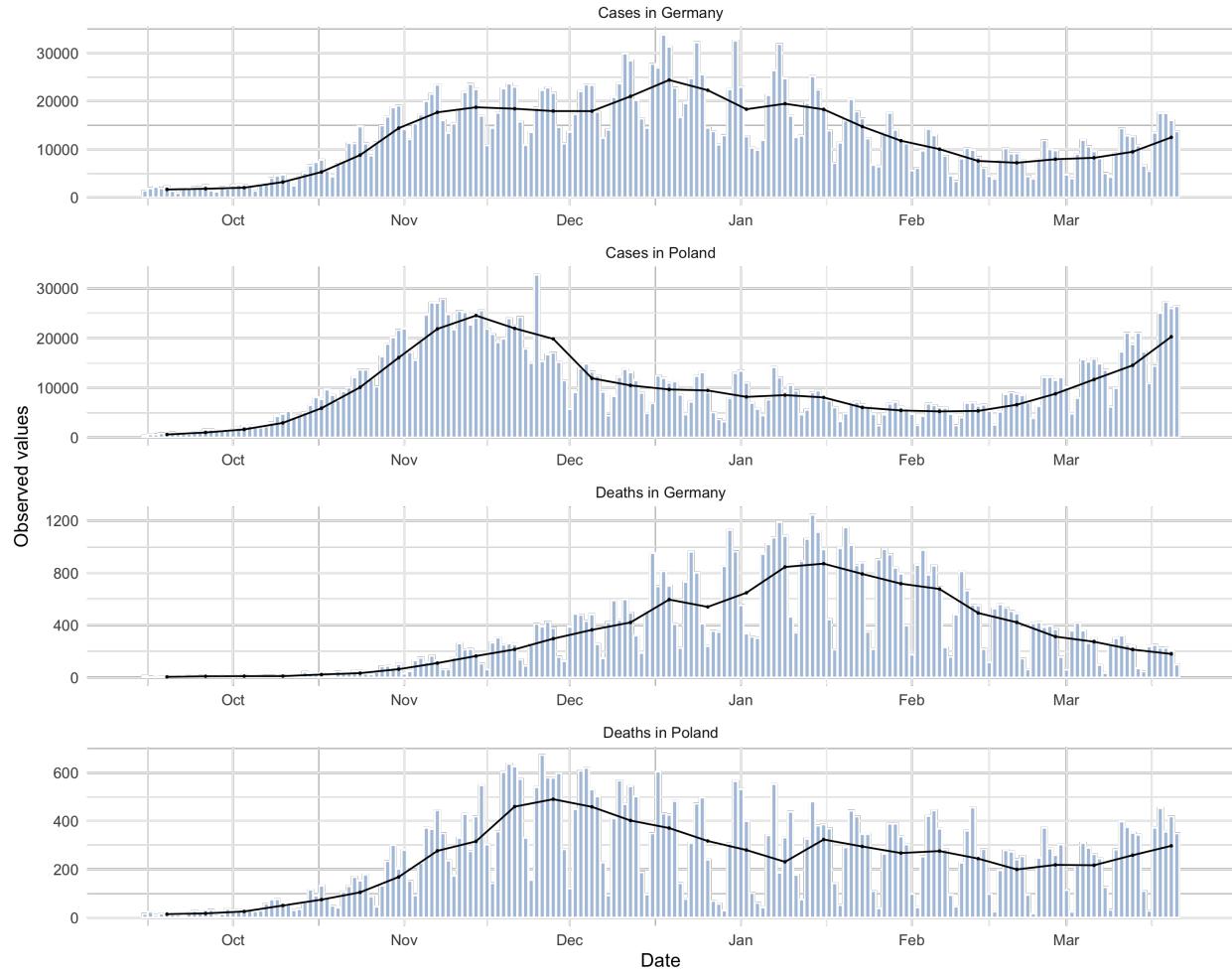


Figure S6: Visualisation of daily report data. The black line represents weekly data divided by seven. Data were last accessed through the German and Polish Forecast Hub on August 21 2021.

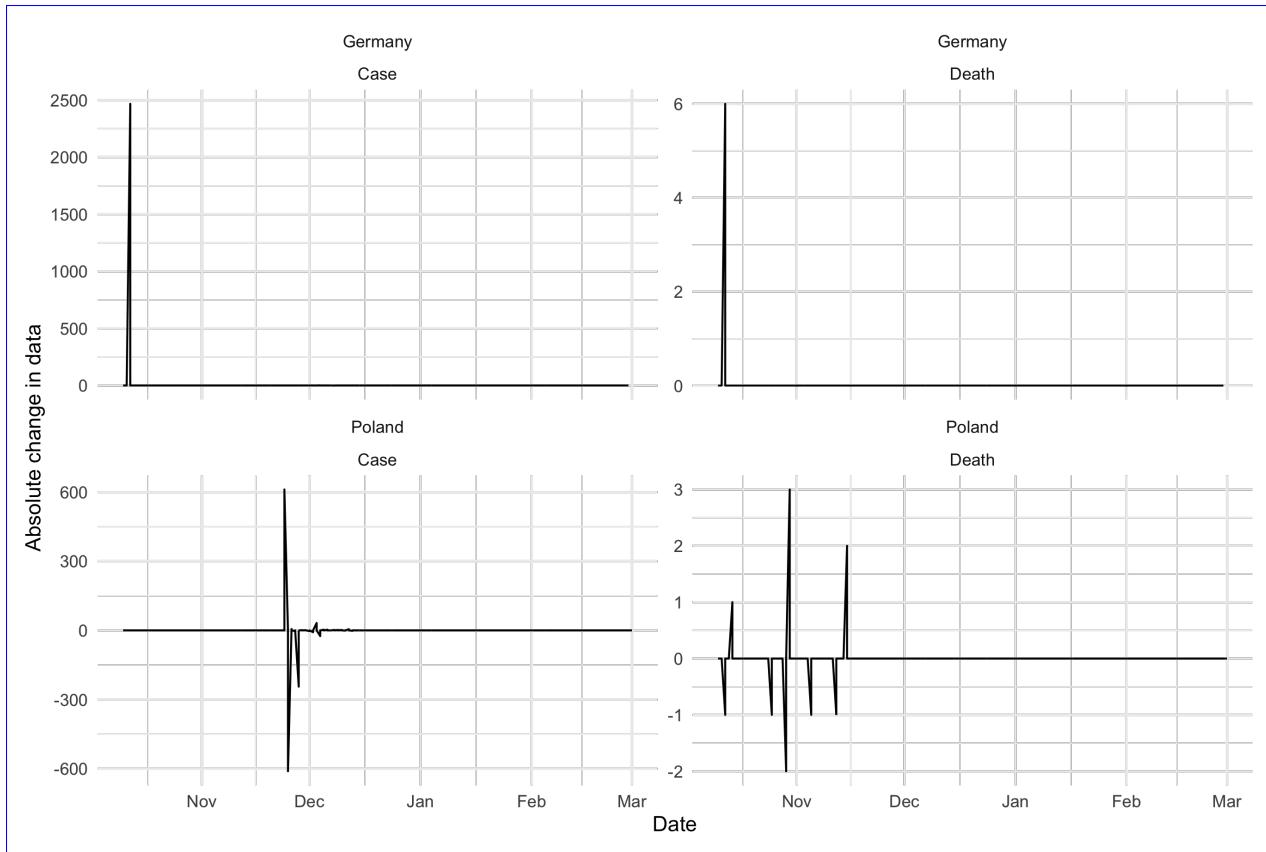


Figure S7: [Visualisation of the absolute difference between the daily report data at the time and the data now](#). In Germany, there were zero cases and deaths reported on 2020-10-12, and only later 2467 cases and 6 deaths were added. Data were last accessed through the German and Polish Forecast Hub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

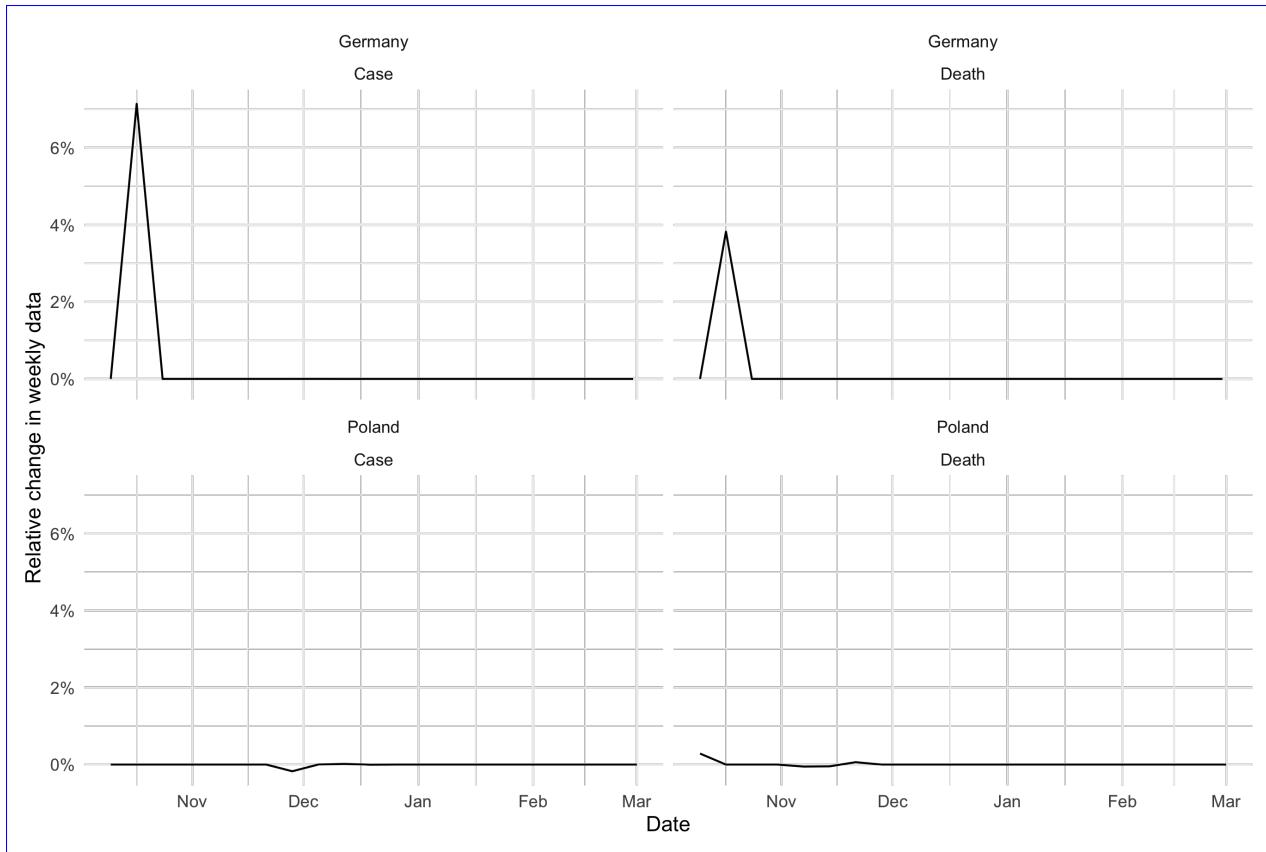


Figure S8: [Visualisation of the relative difference between the weekly report data at the time and the data now](#). Apart from the data that was retrospectively added on 2020-10-12, data updates did not have a noticeable effect on weekly data (as shown in the forecasting application). Data were last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

526 A.8 Visualisation of scores and forecasts 1, 3, 4 weeks ahead

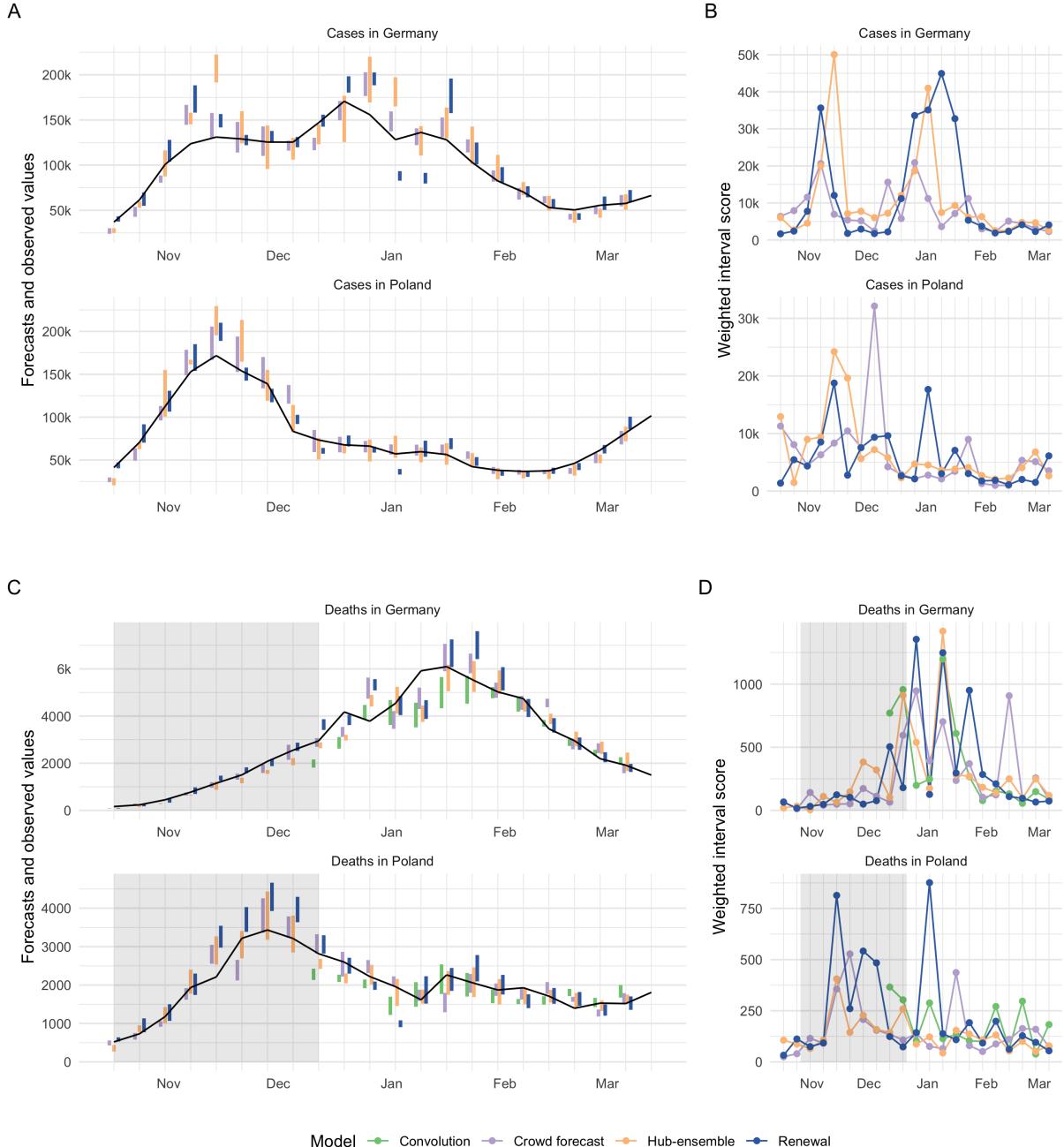


Figure S9: A, C: Visualisation of 50% prediction intervals of one week ahead forecasts against the reported values. Forecasts that were not scored (because there was no complete set of death forecasts available) are greyed out. B, D: Visualisation of corresponding WIS.

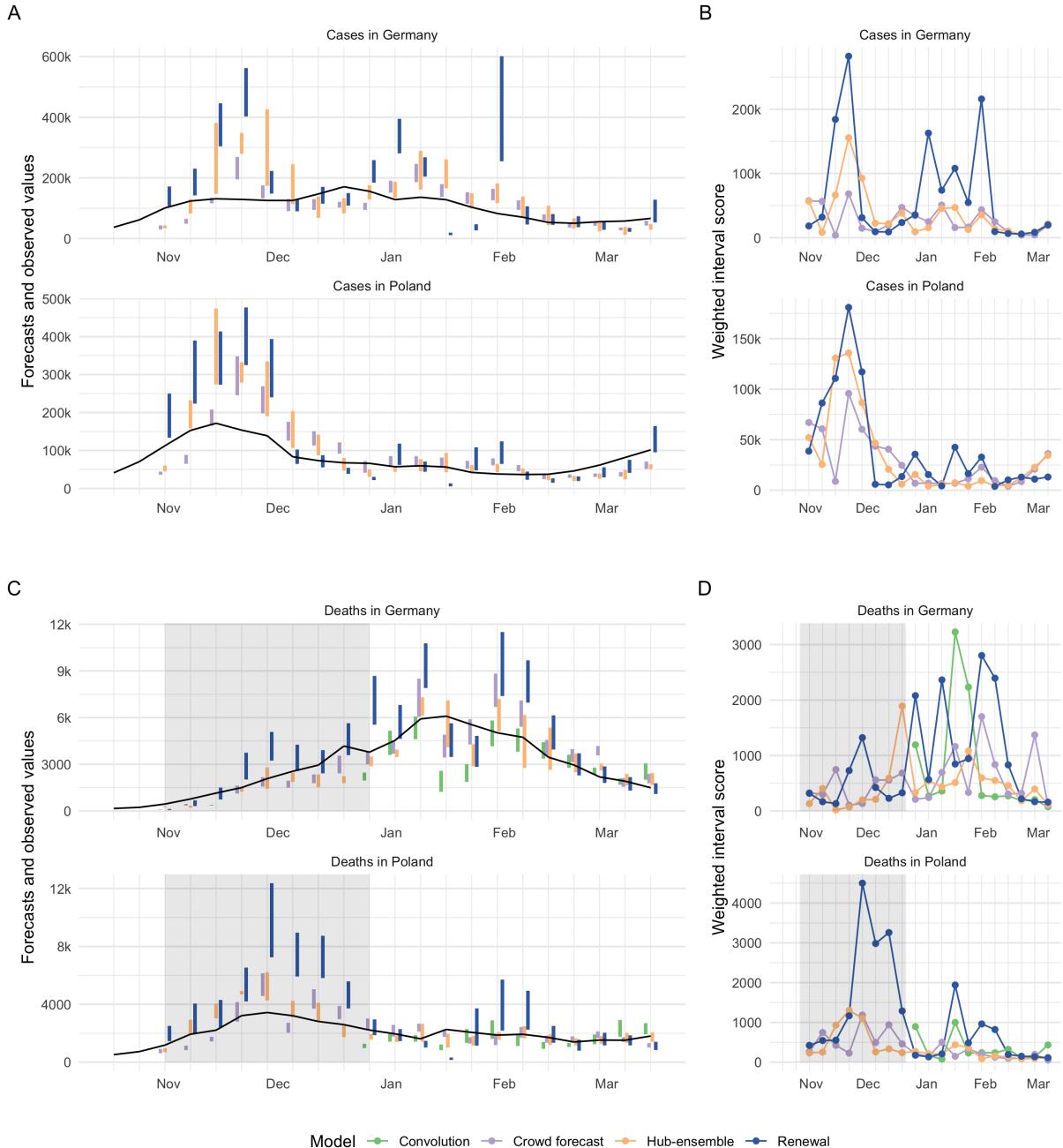


Figure S10: A, C: Visualisation of 50% prediction intervals of three week ahead forecasts against the reported values. Forecasts that were not scored (because there was no complete set of death forecasts available) are greyed out. B, D: Visualisation of corresponding WIS.

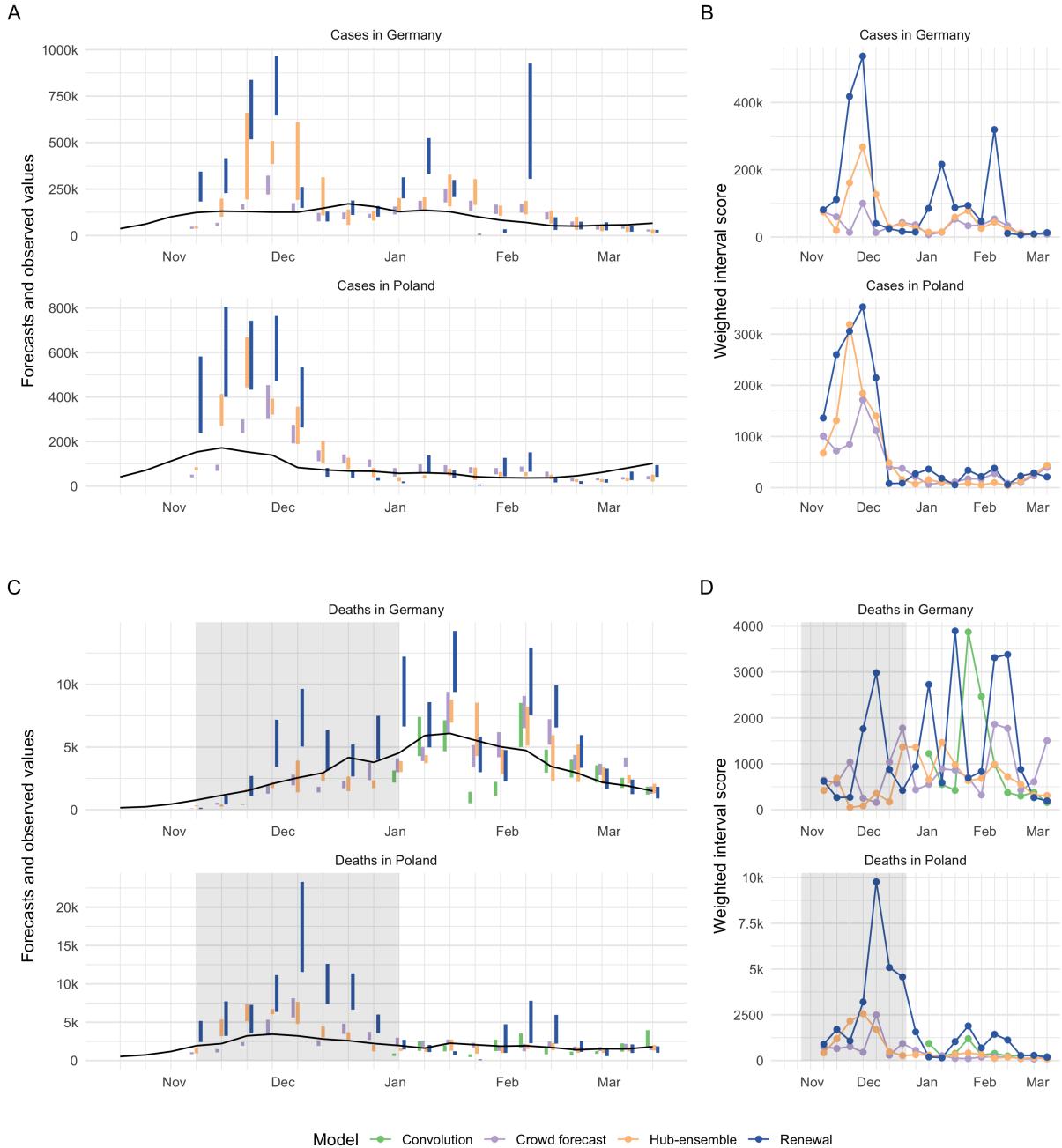


Figure S11: A, C: Visualisation of 50% prediction intervals of four week ahead forecasts against the reported values. Forecasts that were not scored (because there was no complete set of death forecasts available) are greyed out. B, D: Visualisation of corresponding WIS.

527 **A.9 Distribution of scores**

528 **A.9.1 Absolute scores**

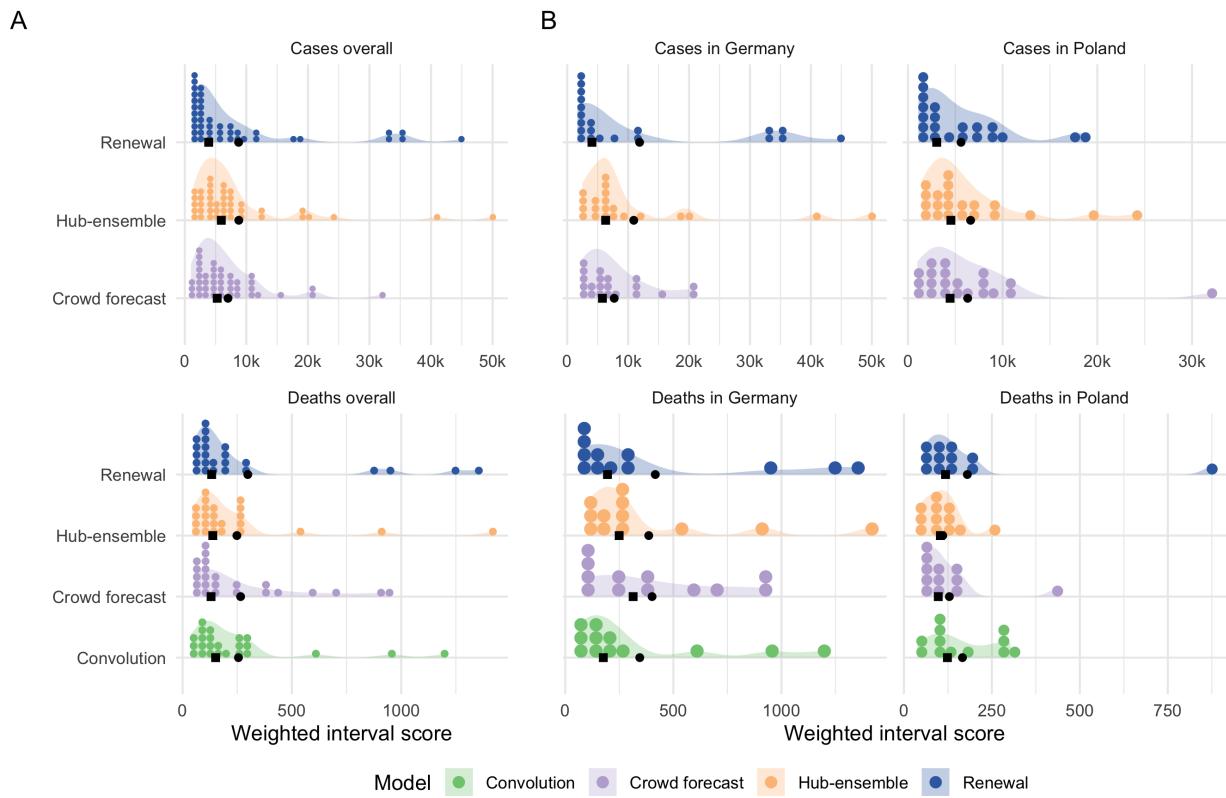


Figure S12: A: Distribution of weighted interval scores for one week ahead forecasts of the different models and forecast targets. B: Distribution of WIS separate by country.

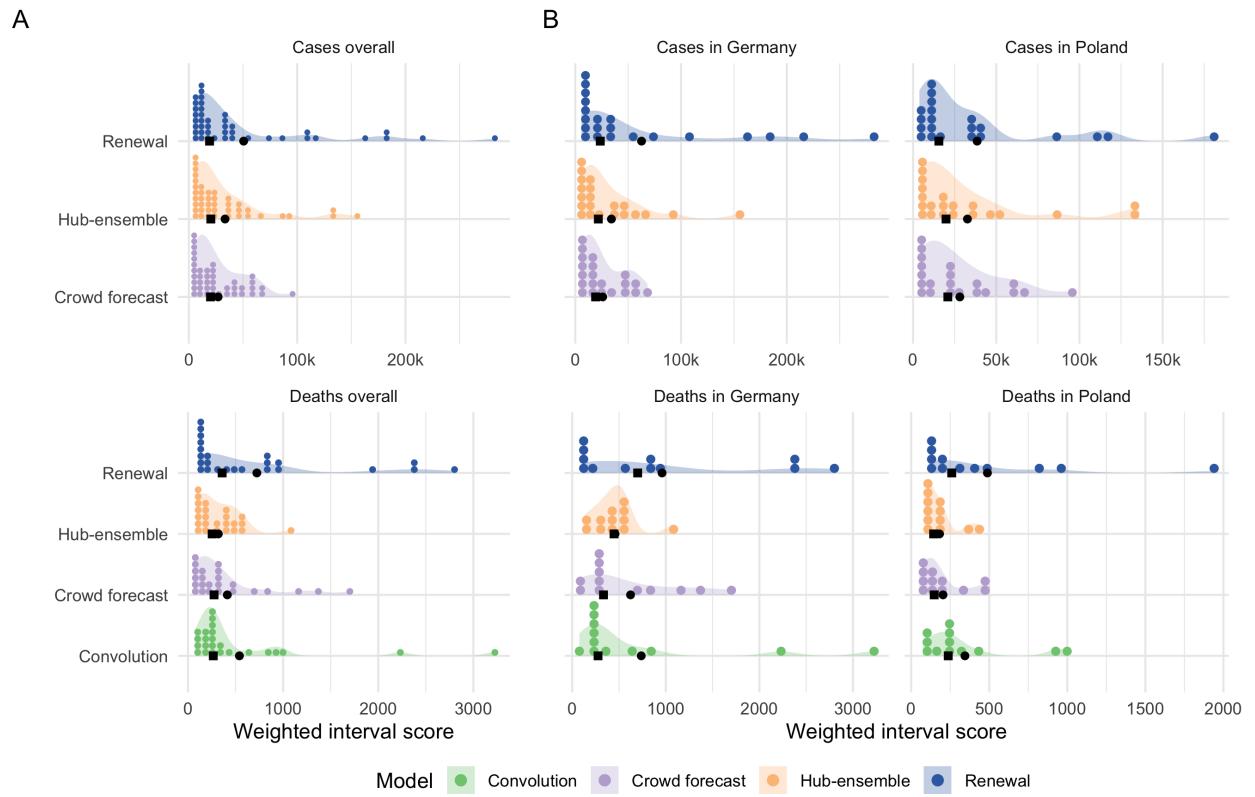


Figure S13: A: Distribution of weighted interval scores for three week ahead forecasts of the different models and forecast targets. B: Distribution of WIS separate by country.

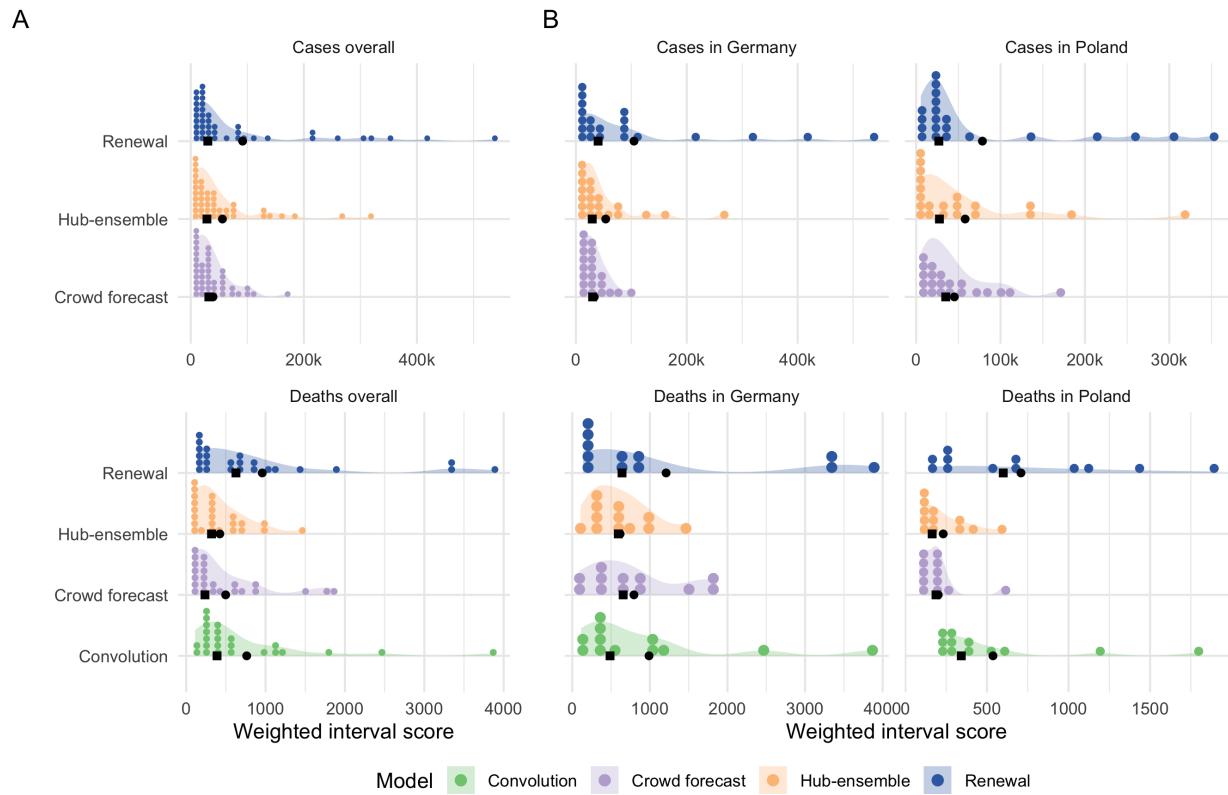


Figure S14: A: Distribution of weighted interval scores for four week ahead forecasts of the different models and forecast targets. B: Distribution of WIS separate by country.

529 **A.9.2 Ranks achieved by forecasts**

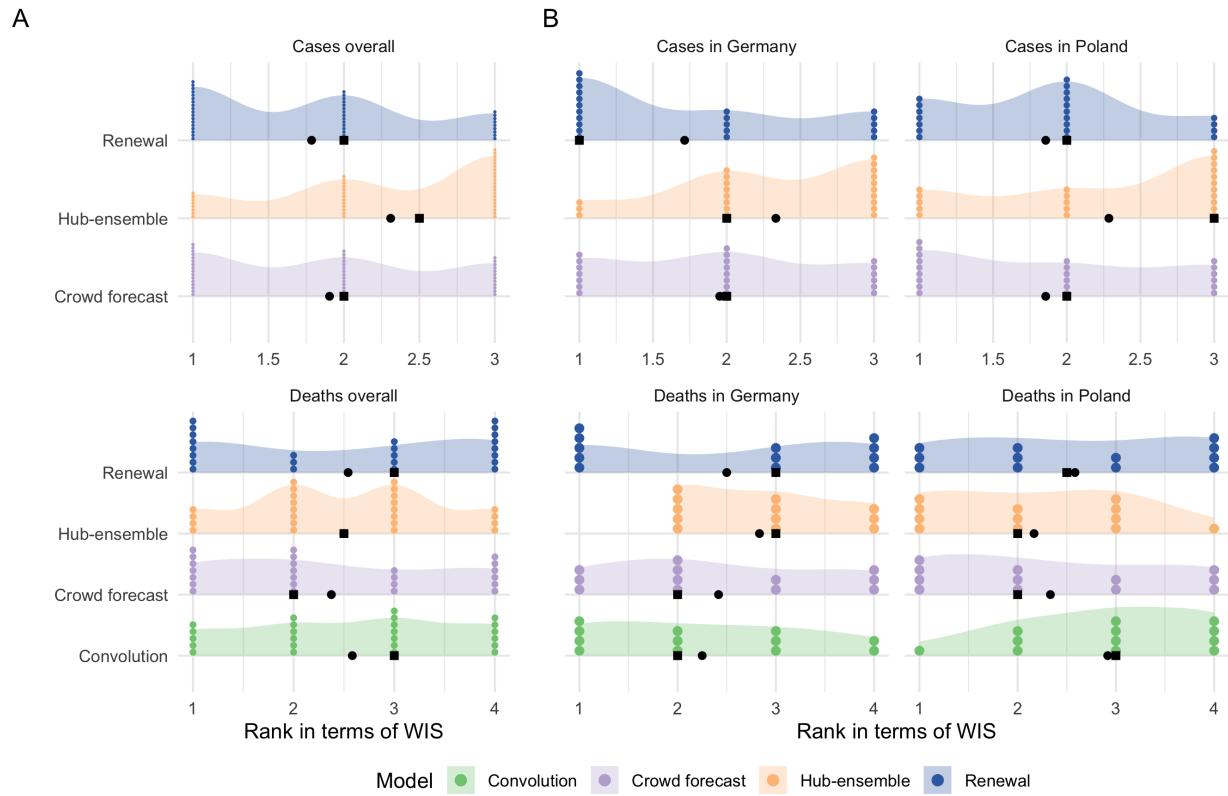


Figure S15: A: Distribution of the ranks (determined by the weighted interval score) for one week ahead forecasts of the different models and forecast targets. B: Distribution of ranks separate by country.

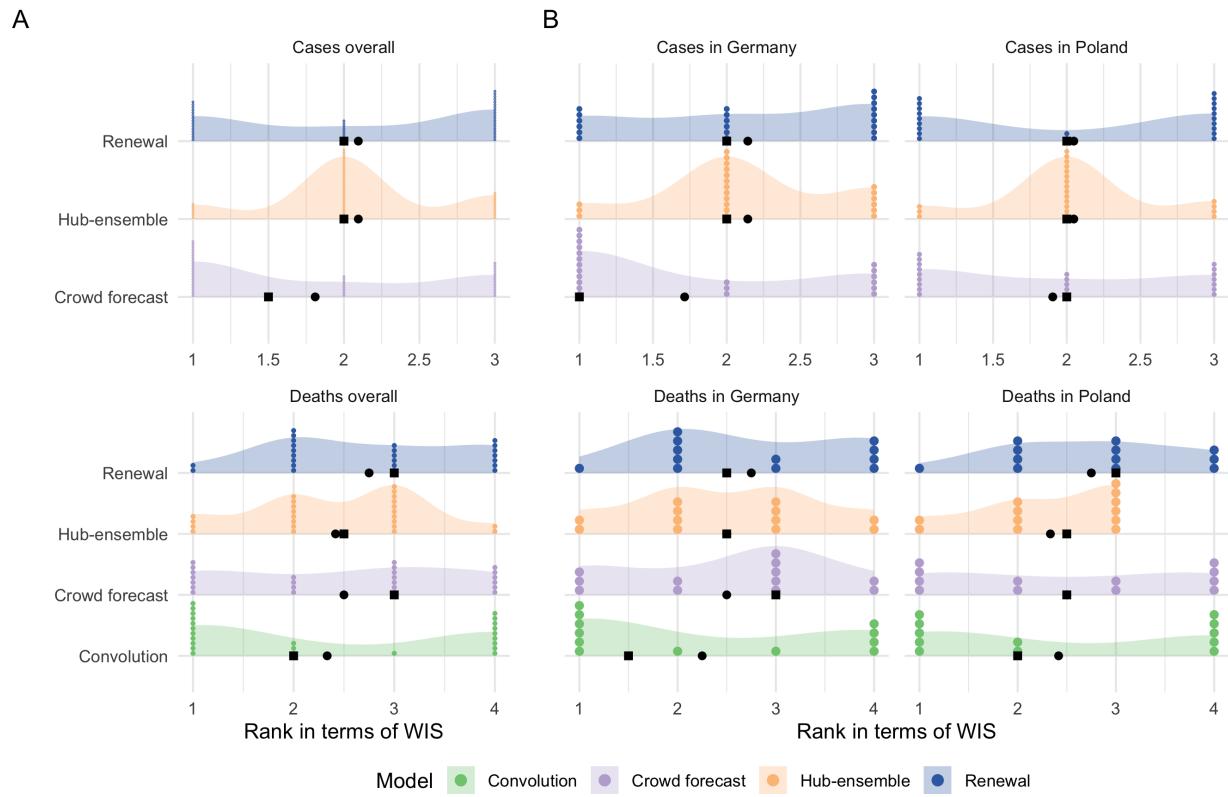


Figure S16: A: Distribution of the ranks (determined by the weighted interval score) for two week ahead forecasts of the different models and forecast targets. B: Distribution of ranks separate by country.

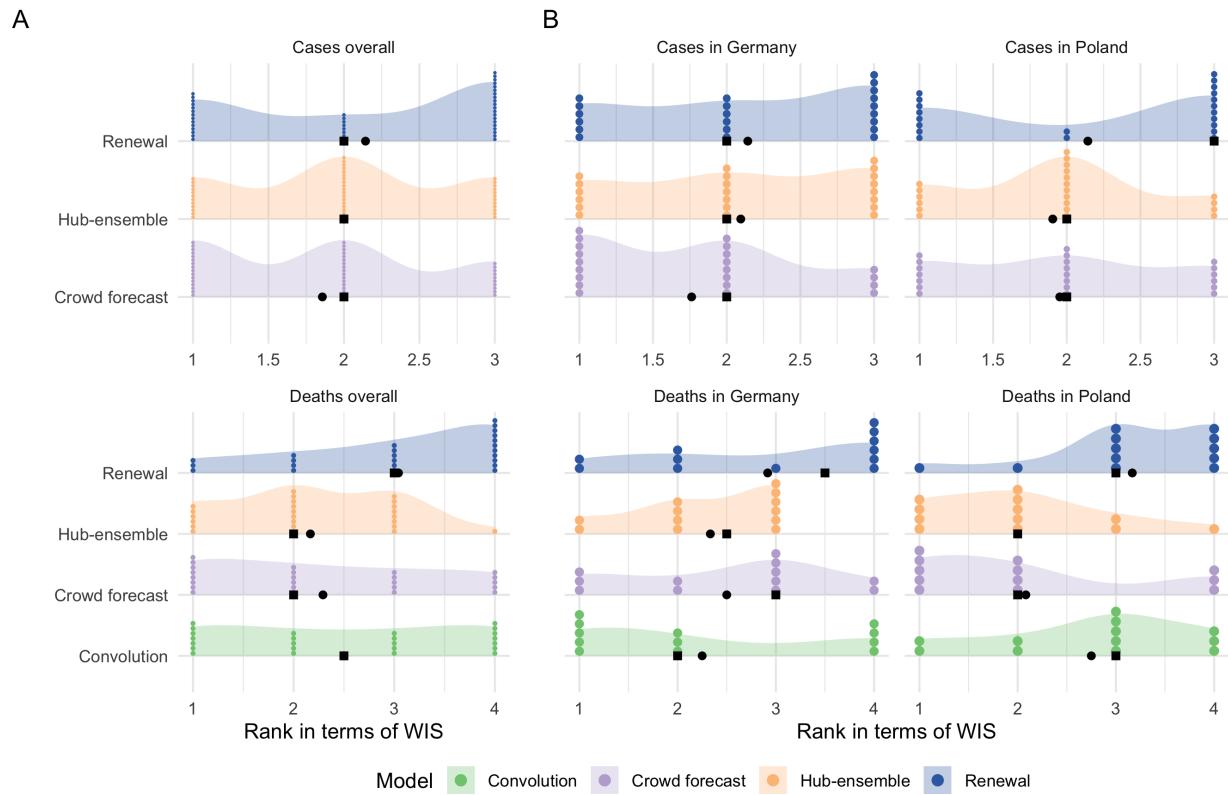


Figure S17: A: Distribution of the ranks (determined by the weighted interval score) for three week ahead forecasts of the different models and forecast targets. B: Distribution of ranks separate by country.

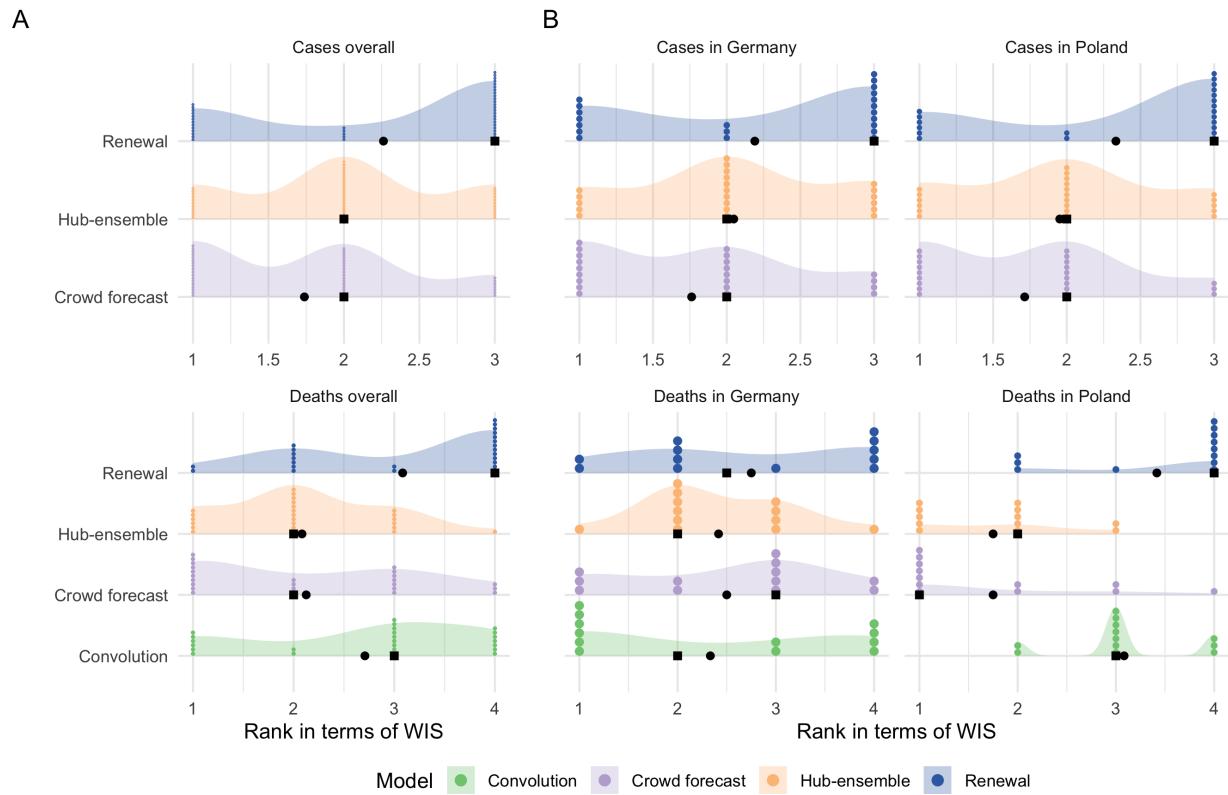


Figure S18: A: Distribution of the ranks (determined by the weighted interval score) for four week ahead forecasts of the different models and forecast targets. B: Distribution of ranks separate by country.

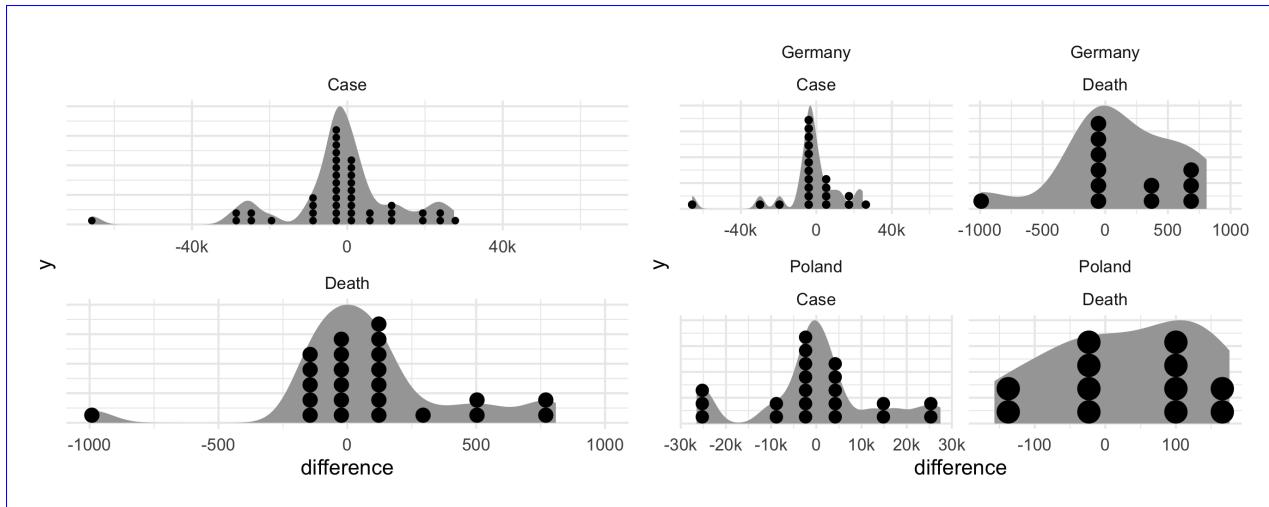


Figure S19: Density plot with the difference in WIS between the Crowd forecast and the Hub ensemble (values below zero mean better performance of the Crowd forecasts) for a 2 week ahead forecast horizon.

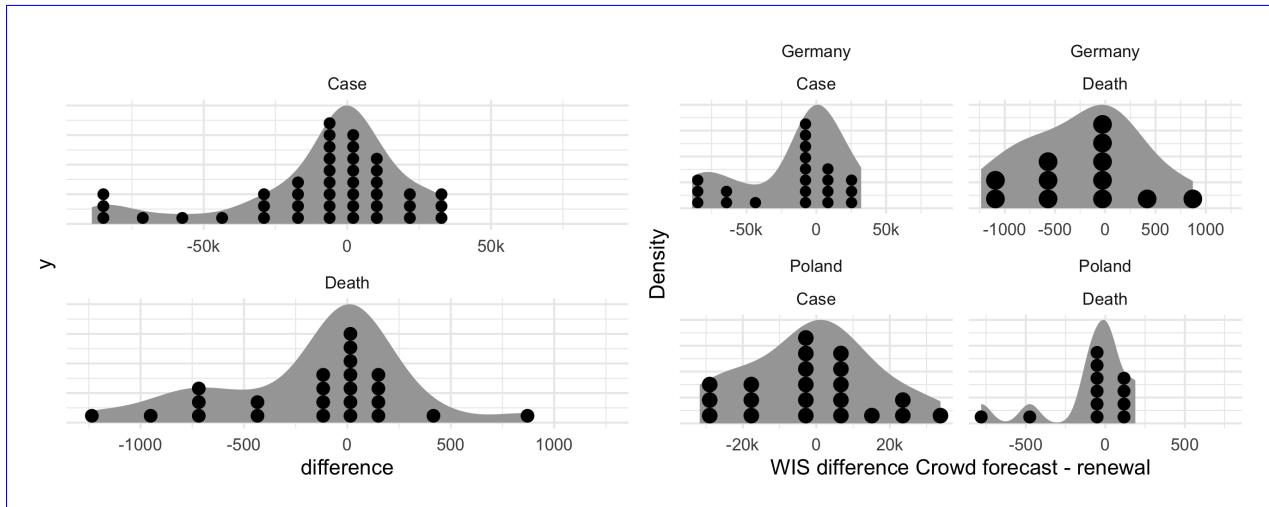


Figure S20: Density plot with the difference in WIS between the Crowd forecast and the Renewal model (values below zero mean better performance of the Crowd forecasts) for a 2 week ahead forecast horizon.

530 **A.10 Comparison of ensembles**

531 **A.10.1 Performance visualisation mean ensemble**

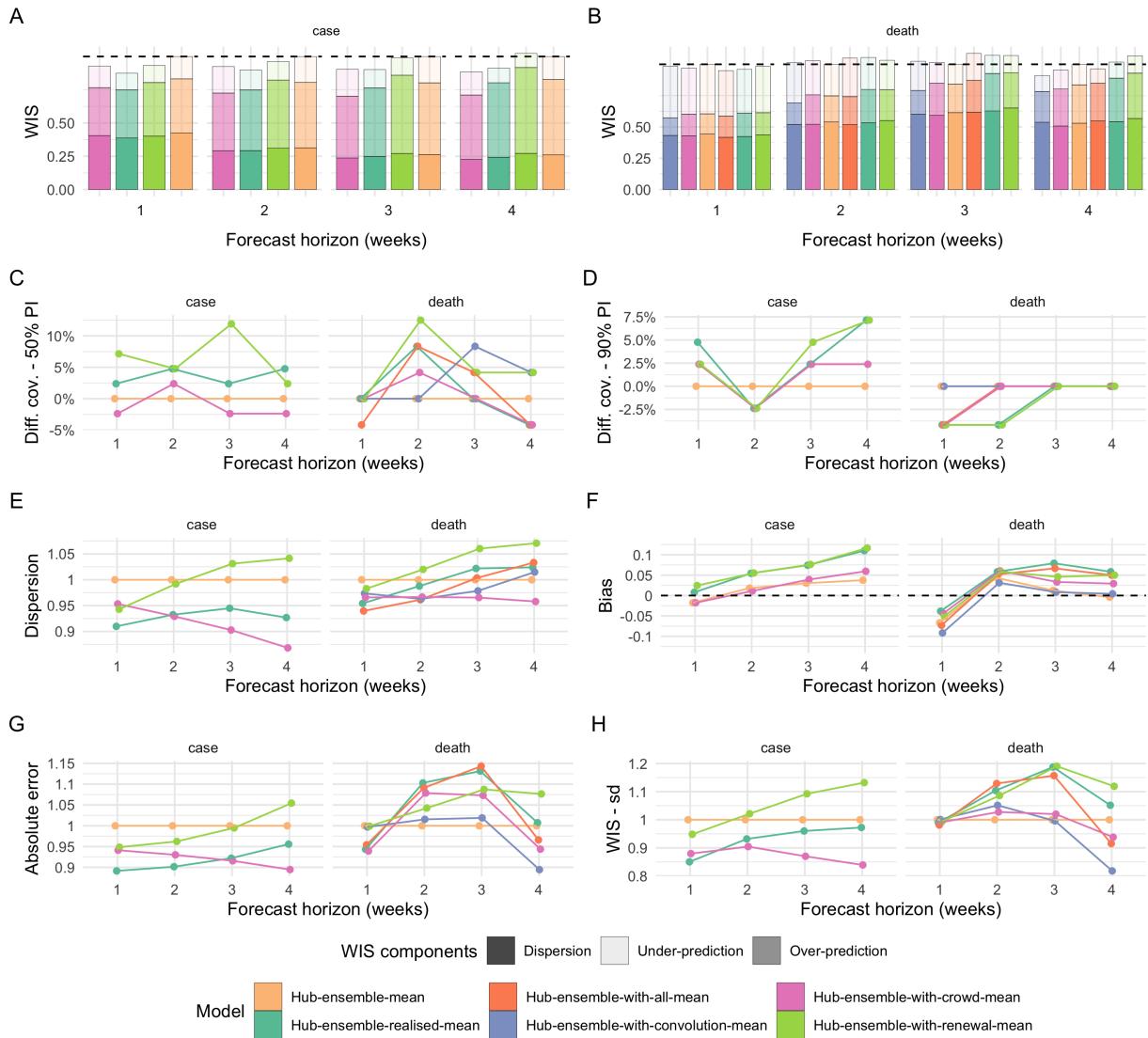


Figure S21: Visualisation of aggregate performance metrics across forecast horizons for the different versions of the Hub mean ensemble. “Hub-ensemble” ~~excludes~~ all our models, Hub-ensemble-all ~~includes~~ all of our models, “Hub-ensemble-real~~Hub-ensemble-realised~~” is the ~~real actual~~ hub-ensemble ~~with observed in reality, which includes~~ the renewal model and the crowd forecasts ~~included, but not the convolution model~~. Values (except for Bias) are computed as differences to the Hub ensemble ~~excluding 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. C, D: Standard deviation to the median forecast Hub ensemble (lower values indicate better performance). C, D: Standard deviation

532 **A.10.2 Tables median 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										
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	
1 wk ahead	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
	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
2 wk ahead	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										
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	
1 wk ahead	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
	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
2 wk ahead	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

533 **A.10.3 Tables mean ensemble**

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

534 **A.11 Sensitivity analysis**

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

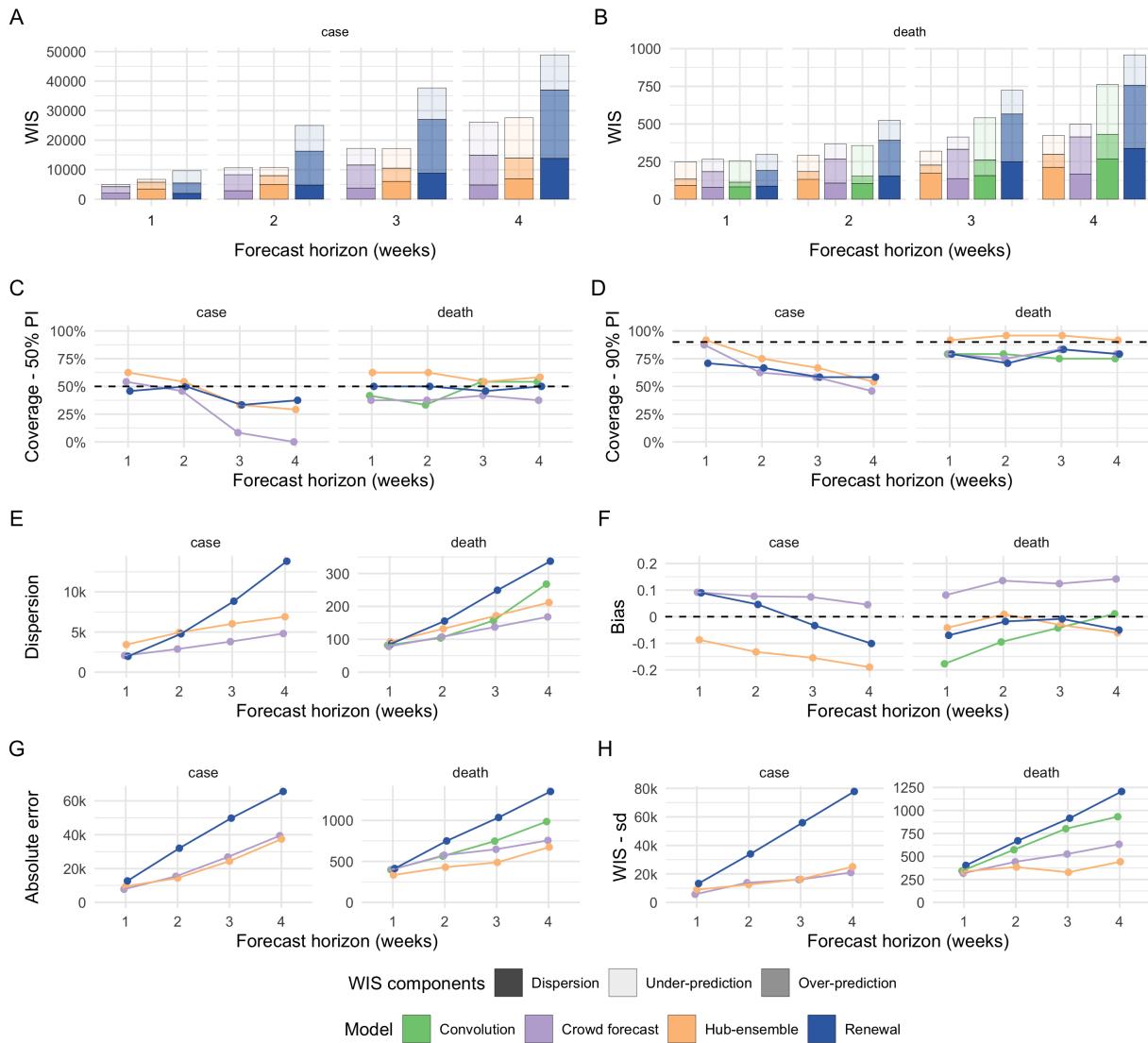


Figure S22: Visualisation of aggregate performance metrics across forecast horizons only for the period from October–December 14th 2020 on where all models were available. A, B: mean weighted interval score (WIS, lower indicates better performance) across horizons. WIS is decomposed into its components 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
2 wk ahead	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
	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
4 wk ahead	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
	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

⁵³⁸ **A.12 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.

Table S10: Overview of the models and ensembles used. (*continued*)

Name	Explanation
Hub-ensemble-with-convolution,	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-convolution-mean	
Hub-ensemble-with-all,	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.
Hub-ensemble-with-all-mean	
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'

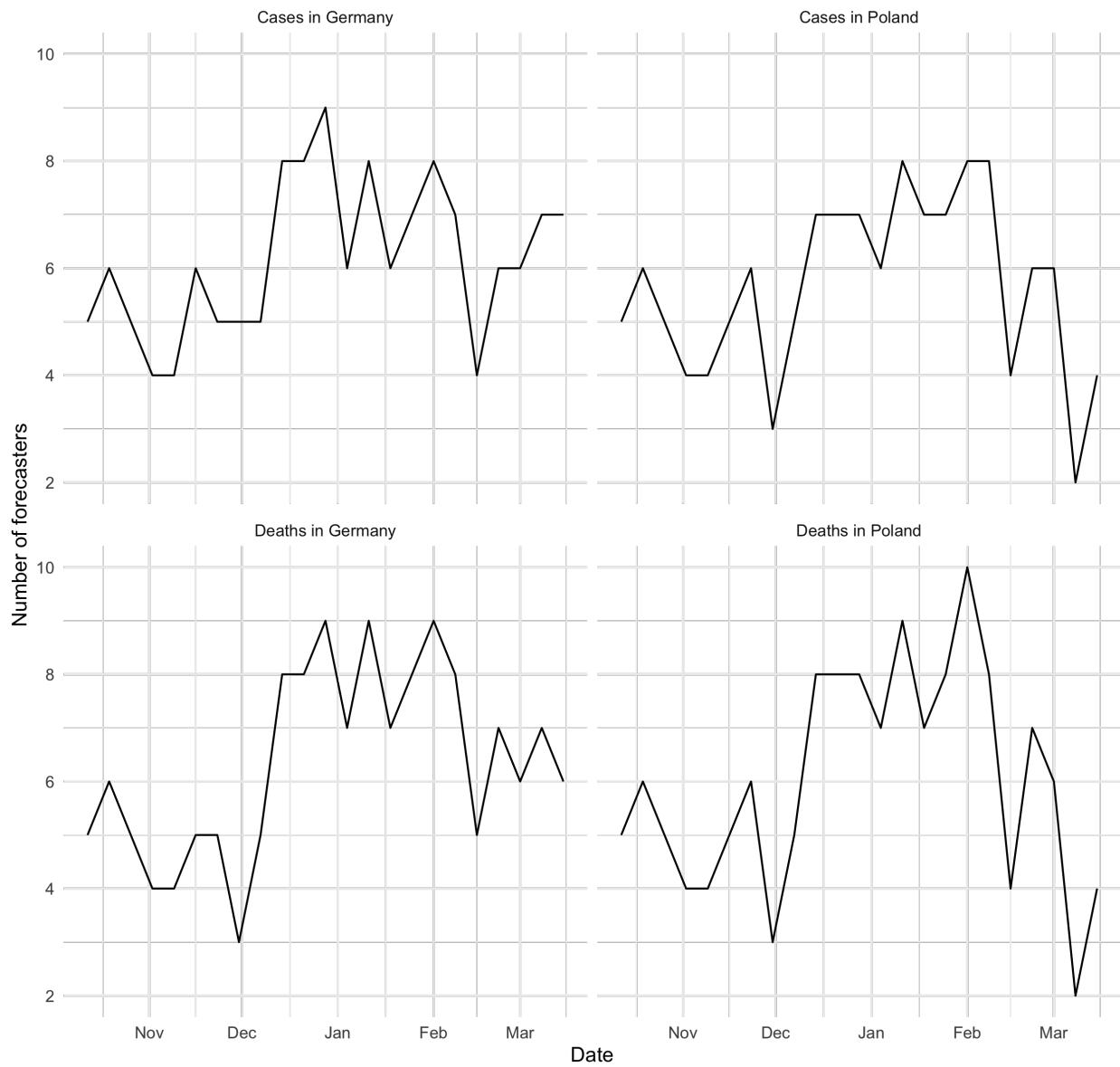


Figure S23: Number of participants who submitted a forecast over time.

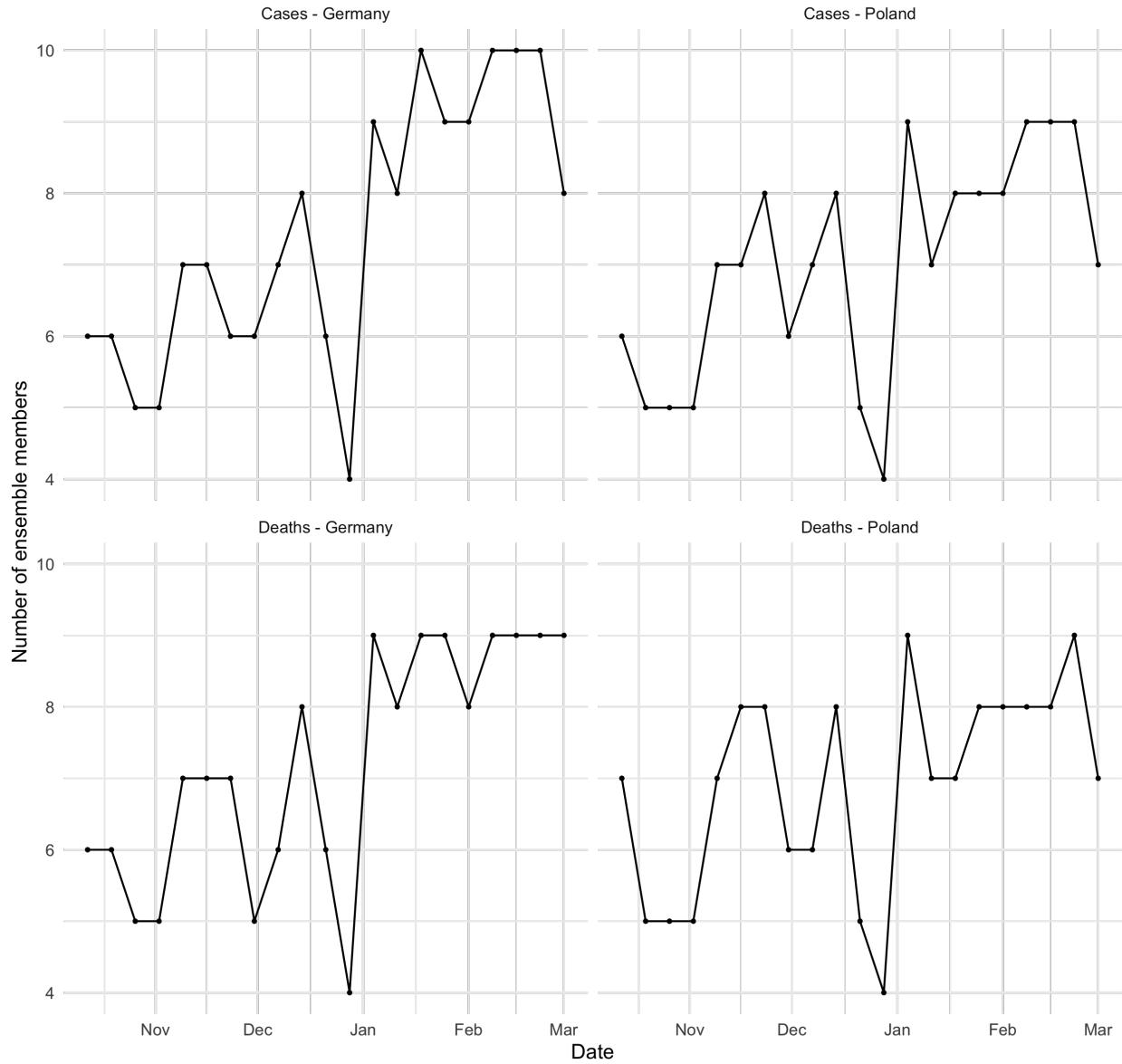


Figure S24: Number of member models (including our crowd forecasts and the renewal model) in the official Hub ensemble. Note that the renewal model was not included in the ensemble on December 28th 2020.

539 A.13 Comparison of crowd forecasts and application baseline

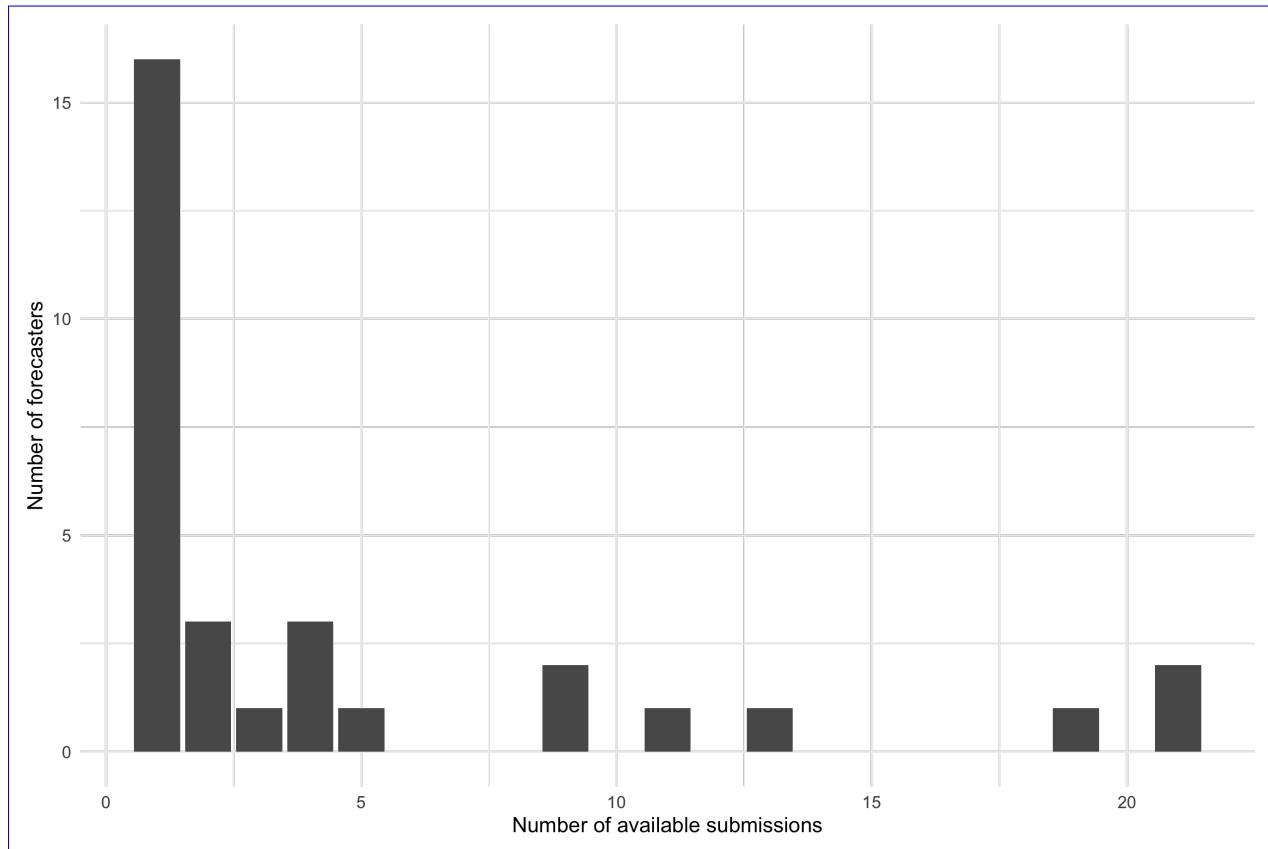


Figure S25: Crowd forecasts and baseline shown in the application for a two week horizon. Shown are the 90% and 50% prediction intervals as well as the median (in order of increasing opacity). For any given point in time, the baseline shown in red is what forecasters saw when they opened the app (the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

- 540 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* [Internet]. 2019 Jan 24 [cited 2021 May 30];9(1, 1):683. Available from: <https://www.nature.com/articles/s41598-018-36361-9>

541

- 542 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 [Internet]. 2019
543 Feb 19 [cited 2021 Oct 13];116(8):3146–54. Available from: <https://www.pnas.org/content/116/8/3146>
- 544 3. Shaman J, Karspeck A. Forecasting seasonal outbreaks of influenza. PNAS [Internet]. 2012 Dec 11
545 [cited 2021 Oct 13];109(50):20425–30. Available from: <https://www.pnas.org/content/109/50/20425>
- 546 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 [Internet]. 2016 Jul 22 [cited 2021 Oct 13];16(1):357. Available from: <https://doi.org/10.1186/s12879-016-1669-x>
- 547
- 548 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 [Internet]. 2019 Nov 26 [cited 2021
549 May 30];116(48):24268–74. Available from: <https://www.pnas.org/content/116/48/24268>
- 550 6. Yamana TK, Kandula S, Shaman J. Superensemble forecasts of dengue outbreaks. Journal of The
Royal Society Interface [Internet]. 2016 Oct 31 [cited 2021 May 30];13(123):20160410. Available from:
551 <https://royalsocietypublishing.org/doi/full/10.1098/rsif.2016.0410>
- 552 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
[Internet]. 2021 Mar 4 [cited 2021 Mar 6];18(3):e1003542. Available from: <https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1003542>
- 553
- 554 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 [Internet]. 2018 Mar 1 [cited 2021 May 30];22:13–21.
555 Available from: <https://www.sciencedirect.com/science/article/pii/S1755436517301275>
- 556 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 [Internet]. 2019 Feb 11 [cited 2019 Sep 16];15(2):e1006785.
557 Available from: <https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006785>

- 558 10. Del Valle SY, McMahon BH, Asher J, Hatchett R, Lega JC, Brown HE, et al. Summary results of the
559 2014-2015 DARPA Chikungunya challenge. *BMC Infectious Diseases* [Internet]. 2018 May 30 [cited
2021 Oct 13];18(1):245. Available from: <https://doi.org/10.1186/s12879-018-3124-7>
- 560 11. Farrow DC, Brooks LC, Hyun S, Tibshirani RJ, Burke DS, Rosenfeld R. A human judgment approach
561 to epidemiological forecasting. *PLOS Computational Biology* [Internet]. 2017 Mar 10 [cited 2021 Jul
29];13(3):e1005248. Available from: <https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005248>
- 562 12. Funk S, Abbott S, Atkins BD, Baguelin M, Baillie JK, Birrell P, et al. Short-term forecasts to inform
563 the response to the Covid-19 epidemic in the UK. *medRxiv* [Internet]. 2020 Nov 13 [cited 2020
Nov 28];2020.11.11.20220962. Available from: <https://www.medrxiv.org/content/10.1101/2020.11.11.20220962v1>
- 564 13. Cramer E, Reich NG, Wang SY, Niemi J, Hannan A, House K, et al. COVID-19 Forecast Hub:
565 4 December 2020 snapshot [Internet]. Zenodo; 2020 [cited 2021 May 29]. Available from: <https://zenodo.org/record/3963371>
- 566 14. Cramer E, Ray EL, Lopez VK, Bracher J, Brennen A, Rivadeneira AJC, et al. Evaluation of individual
567 and ensemble probabilistic forecasts of COVID-19 mortality in the US. *medRxiv* [Internet]. 2021 Feb
5 [cited 2021 Apr 6];2021.02.03.21250974. Available from: <https://www.medrxiv.org/content/10.1101/2021.02.03.21250974v1>
- 568 15. Bracher J, Wolffram D, Deuschel J, Görden K, Ketterer JL, Ullrich A, et al. Short-term forecasting
569 of COVID-19 in Germany and Poland during the second wave – a preregistered study. *medRxiv*
[Internet]. 2021 Jan 11 [cited 2021 Apr 1];2020.12.24.20248826. Available from: <https://www.medrxiv.org/content/10.1101/2020.12.24.20248826v2>
- 570 16. Bracher J, Wolffram D, Deuschel J, Görden K, Ketterer JL, Ullrich A, et al. National and subnational
571 short-term forecasting of COVID-19 in Germany and Poland, early 2021. 2021 Nov 8 [cited 2021
Nov 18];2021.11.05.21265810. Available from: <https://www.medrxiv.org/content/10.1101/2021.11.05.21265810v1>

- 572 17. European Covid-19 Forecast Hub. European Covid-19 Forecast Hub [Internet]. 2021 [cited 2021 May
573 30]. Available from: <https://covid19forecasthub.eu/>
- 574 18. Reich NG, McGowan CJ, Yamana TK, Tushar A, Ray EL, Osthuis D, et al. Accuracy of real-time
multi-model ensemble forecasts for seasonal influenza in the U.S. PLOS Computational Biology
[Internet]. 2019 Nov 22 [cited 2020 Aug 7];15(11):e1007486. Available from: <https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1007486>
575
- 576 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 [Internet]. 2014 Aug 1 [cited
577 2021 May 30];23(4):290–5. Available from: <https://doi.org/10.1177/0963721414534257>
- 578 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 [Internet]. 2016 Apr 22
[cited 2021 May 30];63(3):691–706. Available from: <https://pubsonline.informs.org/doi/abs/10.1287/mnsc.2015.2374>
579
- 580 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 [Internet].
581 2020 Sep 1 [cited 2021 Oct 13];3(3):267–85. Available from: <https://doi.org/10.1177/2515245920919667>
- 582 22. ReplicationMarkets. Replication Markets – Reliable research replicates... you can bet on it. [Internet].
583 2020 [cited 2021 Oct 13]. Available from: <https://www.replicationmarkets.com/>
- 584 23. Servan-Schreiber E, Wolfers J, Pennock DM, Galebach B. Prediction Markets: Does Money Mat-
ter? Electronic Markets [Internet]. 2004 Sep 1 [cited 2021 Oct 13];14(3):243–51. Available
from: <http://www.informaworld.com/openurl?genre=article&doi=10.1080/1019678042000245254&magic=crossref%7C%7CD404A21C5BB053405B1A640AFFD44AE3>
585
- 586 24. McAndrew TC, Reich NG. An expert judgment model to predict early stages of the COVID-19 outbreak
in the United States. medRxiv [Internet]. 2020 Sep 23 [cited 2020 Sep 23];2020.09.21.20196725.
587 Available from: <https://www.medrxiv.org/content/10.1101/2020.09.21.20196725v1>

- 588 25. Recchia G, Freeman ALJ, Spiegelhalter D. How well did experts and laypeople forecast the size of the
COVID-19 pandemic? PLOS ONE [Internet]. 2021 May 5 [cited 2021 Jun 2];16(5):e0250935. Available
from: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0250935>
- 589
- 590 26. Metaculus. A Preliminary Look at Metaculus and Expert Forecasts [Internet]. 2020 [cited 2021 May
30]. Available from: <https://www.metaculus.com/news/2020/06/02/LRT/>
- 591
- 592 27. Hypermind. Hypermind | Supercollective intelligence for decision makers [Internet]. Hypermind; 2021
[cited 2021 Oct 13]. Available from: <https://www.hypermind.com/en/>
- 593
- 594 28. CSET Foretell. CSET Foretell [Internet]. 2021 [cited 2021 Oct 13]. Available from: <https://www.cset-foretell.com/>
- 595
- 596 29. PredictIt. PredictIt [Internet]. 2021 [cited 2021 Oct 13]. Available from: <https://www.predictit.org/>
- 597
- 598 30. Held L, Meyer S, Bracher J. Probabilistic forecasting in infectious disease epidemiology: The 13th
Armitage lecture. Statistics in Medicine [Internet]. 2017 [cited 2019 Sep 16];36(22):3443–60. Available
from: <https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.7363>
- 599
- 600 31. Bosse NI, Abbott S, EpiForecasts, Funk S. Crowdforecastr: Eliciting crowd forecasts in r shiny. 2020.
- 601
- 602 32. Bosse NI, Abbott S, EpiForecasts, Funk S. Covid.german.forecasts: Forecasting covid-19 related
metrics for the german/poland forecast hub. 2020.
- 603
- 604 33. ECDC. Download historical data (to 14 December 2020) on the daily number of new reported COVID-
19 cases and deaths worldwide [Internet]. European Centre for Disease Prevention and Control; 2020
[cited 2021 May 30]. Available from: <https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide>
- 605
- 606 34. RKI. RKI - Coronavirus SARS-CoV-2 - Aktueller Lage-/Situationsbericht des RKI zu COVID-19
[Internet]. 2021 [cited 2021 May 30]. Available from: https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Situationsberichte/Gesamt.html
- 607

- 608 35. Forsal.pl. Rozbieżności w statystykach koronawirusa. 22 tys. przypadków będą doliczone do ogólnej liczby wyników [Internet]. 2020 [cited 2021 May 30]. Available from: <https://forsal.pl/lifestyle/zdrowie/artykuly/8017628,rozbieznosci-w-statystykach-koronawirusa-22-tys-przypadkow-beda-doliczone-do-ogolnej-liczby-wynikow.html>
- 609
- 610 36. Ärzteblatt DÄG Redaktion Deutsches. SARS-CoV-2-Diagnostik: RKI passt Testempfehlungen an [Internet]. Deutsches Ärzteblatt; 2020 [cited 2021 May 30]. Available from: <https://www.aerzteblatt.de/nachrichten/118001/SARS-CoV-2-Diagnostik-RKI-passt-Testempfehlungen-an>
- 611
- 612 37. Fay C, Guyader V, Rochette S, Girard C. Golem: A framework for robust shiny applications [Internet]. 2021. Available from: <https://github.com/ThinkR-open/golem>
- 613
- 614 38. Chang W, Cheng J, Allaire J, Sievert C, Schloerke B, Xie Y, et al. Shiny: Web application framework for r [Internet]. 2021. Available from: <https://CRAN.R-project.org/package=shiny>
- 615
- 616 39. Our World in Data. COVID-19 Data Explorer [Internet]. Our World in Data; 2020 [cited 2021 May 30]. Available from: <https://ourworldindata.org/coronavirus-data-explorer>
- 617
- 618 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;-(-):-.
- 619
- 620 41. Fraser C. Estimating Individual and Household Reproduction Numbers in an Emerging Epidemic. PLOS ONE [Internet]. 2007 Aug 22 [cited 2021 Sep 29];2(8):e758. Available from: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0000758>
- 621
- 622 42. epiforecasts.io/covid. Covid-19: Temporal variation in transmission during the COVID-19 outbreak [Internet]. Covid-19; 2020 [cited 2021 May 30]. Available from: <https://epiforecasts.io/covid/>
- 623
- 624 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. 2021 Mar 18 [cited 2021 Oct 14];2020.10.18.20214585. Available from: <https://www.medrxiv.org/content/10.1101/2020.10.18.20214585v2>
- 625

- 626 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. 2020 Jun 1
627 [cited 2021 Oct 14];(5:112). Available from: <https://wellcomeopenresearch.org/articles/5-112>
- 628 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 [Internet]. 2020. Available
from: <http://virological.org/t/epidemiological-data-from-the-ncov-2019-outbreak-early-descriptions->
629 from-publicly-available-data/337
- 630 46. Stan Development Team. RStan: The r interface to stan [Internet]. 2020. Available from: <http://mc->
631 stan.org/
- 632 47. Bracher J, Ray EL, Gneiting T, Reich NG. Evaluating epidemic forecasts in an interval format. PLoS
633 Comput Biol. 2021 Feb;17(2):e1008618.
- 634 48. Gneiting T, Raftery AE. Strictly proper scoring rules, prediction, and estimation. Journal of the
American Statistical Association [Internet]. 2007 Mar [cited 2020 Mar 22];102(477):359–78. Available
635 from: <http://www.tandfonline.com/doi/abs/10.1198/016214506000001437>
- 636 49. Bosse NI, Abbott S, EpiForecasts, Funk S. Scoringutils: Utilities for scoring and assessing predictions.
637 2020.
- 638 50. Deutsche Welle. Coronavirus: Germany to impose one-month partial lockdown | DW | 28.10.2020
[Internet]. 2020 [cited 2021 Jun 29]. Available from: <https://www.dw.com/en/coronavirus-germany-to-impose-one-month-partial-lockdown/a-55421241>
639
- 640 51. Ganyani T, Kremer C, Chen D, Torneri A, Faes C, Wallinga J, et al. Estimating the generation interval
for coronavirus disease (COVID-19) based on symptom onset data, march 2020. Eurosurveillance.
641 2020;25(17).
- 642 52. Lauer SA, Grantz KH, Bi Q, Jones FK, Zheng Q, Meredith HR, et al. The incubation period
of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: Estimation and
643 application. Annals of Internal Medicine. 2020;172(9):577–82.

- 644 53. Abbott S, Sherratt K, Bevan J, Gibbs H, Hellewell J, Munday J, et al. Covidregionaldata: Subnational
645 data for the covid-19 outbreak. -. 2020;-(-):-.
- 646 54. Evaluating the use of the reproduction number as an epidemiological tool, using spatio-temporal
trends of the Covid-19 outbreak in England | medRxiv [Internet]. [cited 2021 May 30]. Available from:
647 <https://www.medrxiv.org/content/10.1101/2020.10.18.20214585v1>
- 648 55. Riutort-Mayol G, Bürkner P-C, Andersen MR, Solin A, Vehtari A. Practical hilbert space approximate
bayesian gaussian processes for probabilistic programming [Internet]. 2020. Available from: <https://arxiv.org/abs/2004.11408>
- 649 56. Scott JA, Gandy A, Mishra S, Unwin J, Flaxman S, Bhatt S. Epidemia: Modeling of epidemics using
hierarchical bayesian models [Internet]. 2020. Available from: <https://imperialcollegeLondon.github.io/epidemia/>
- 650 57. Bhatt S, Ferguson N, Flaxman S, Gandy A, Mishra S, Scott JA. Semi-Mechanistic Bayesian modeling
of COVID-19 with Renewal Processes. :14.
- 651
- 652
- 653