Evaluating crowd sourced forecasts of Covid-19 against epidemiological model forecasts in Germany and Poland

Evaluating crowd sourced forecasts of Covid-19 against epidemiological model forecasts in Germany and Poland

Nikos Bosse, Sam Abbott, and Sebastian Funk

Target journal: elife

Abstract

Background Model based forecasts have played an important role in shaping public policy throughout the Covid-19 pandemic. The models, in turn, have been tweaked and shaped by human judgement. Any model forecast is therefore a mix between the researcher's subjective opinion and mechanistic model assumptions.

Methods To discern these two components we looked at forecasts Covid-19 case and death numbers submitted to the German and Polish Forecast Hub between October 2020 until February 2021. Crowd sourced human forecasts were compared against predictions by an untuned epidemiological baseline model from the same working group, as well as against the ensemble of all predictions submitted to the hub.

Results Human forecasts outperformed models (including the Forecast Hub ensemble) on case forecasts, but not on death forecasts. They performed best over longer time horizons and around Christmas where reporting artifacts had a major influence on the data.

Conclusions Expert opinion outperforms models at predicting long-term trends and when dealing with data anomalies, as humans can make use of information, e.g. about potential future policy interventions, not directly available to models. Computer models, however, have an edge in situations like XY and

when modelling epidemiological parameters like the delay between current case numbers and future deaths.

Introduction

The COVID-19 pandemic has resulted in an increase of interest in infectious disease forecasting, and the evaluation of these forecasts. Single model forecasts (Ferguson et al. 2020; IHME COVID-19 health service utilization forecasting team and Murray 2020) were impactful on policy decisions early in the pandemic despite previous work having shown that relying on a single model can lead to less accurate forecasts than decisions based on multiple approaches (Yamana, Kandula, and Shaman 2016; Gneiting and Raftery 2005). Since then several collaborations have sort to evaluate Covid-19 forecasts in the United Kingdom (S. Funk et al. 2020), in the United States of America (E. Cramer et al. 2020; E. Y. Cramer et al. 2021), and in Germany and Poland (J. Bracher et al. 2021). Whilst all of these efforts have successfully delivered more accurate forecasts to policy makers compared to individual forecasters efforts they have struggled to unpick what leads to good Covid-19 forecasts (E. Y. Cramer et al. 2021; J. Bracher et al. 2021; S. Funk et al. 2020).

This has been partly driven by the complexity of the models used to produce the constituent forecasts but also because of the level of expert intervention in most forecasting methods over time, and in response to changes in the pandemic. These issues can be decoupled by separating infectious disease forecasting into "automatic" model derived forecasts and human elicitation forecasts (from now on referred to as crowd forecasts). Model based forecasts have a rich history and have been growing in popularity over the last decade (McGowan et al. 2019; Johansson et al. 2019; Viboud et al. 2018; Sebastian Funk et al. 2019). However, it is unusual for "automated" real-time forecasts (as opposed to retrospective forecasts) to be evaluated with forecasts usually being submitted by individual researchers and therefore liable to change over time in response to percieved performance, changes in the underlying infectious disease processes or for other reasons. A variety of human expert elicitation as well as crowd forecasting projects exist (McAndrew et al. 2021; "A Preliminary Look at Metaculus and Expert Forecasts" n.d.; Tetlock et al. 2014; Atanasov et al. 2016). However, these forecasts usually follow a different format than the ones provided by traditional forecasting models or take on different questions. Unlike these projects the crowd forecasts we develop here have been specifically designed to be comparable to model based forecasts.

In this work, we evaluate two contrasting forecasting approaches that simplify and synthesise some these themes. The first of these approaches is a crowd forecast, where expert and non-expert opinion is combined into a single forecast of cases and deaths in target locations. This represents modellers interventions in forecasts but in a model agnostic format. In the second approach, we use two recently developed short term forecasting methods that make minimal epidemiological assumptions of how notifications are generated over time coupled

with a robust observation model. These models were then not tuned throughout the submission period in order to make a comparison to opinion derived forecasts possible. All of these forecasts were submitted to the German/Poland forecast hub over 21 weeks from the 12th October 2020 to March 1st 2021 and combined, along with other forecasts, into an ensemble used by policy makers as well as being independently evaluated by the research group running the German/Poland forecasting hub.

Methods

Data sources

Data on test positive cases and deaths linked to Covid-19 were provided by the organisers of the German and Polish forecast hub (P/L hub) (J. Bracher et al. 2021). Until December 14th 2020 these data were sourced from the European Centre for Disease Control (ECDC) ("Download Historical Data (to 14 December 2020) on the Daily Number of New Reported COVID-19 Cases and Deaths Worldwide" 2020). After ECDC stopped publishing daily data, observations were sourced from the Robert Koch Institute (RKI) for the remainder of the submission period ("RKI - Coronavirus SARS-CoV-2 - Aktueller Lage-/Situationsbericht Des RKI Zu COVID-19" n.d.). These data are subject to reporting artefacts (such as a retrospective case reporting in Poland on the 24th November "Rozbieżności w statystykach koronawirusa. 22 tys. przypadków będą doliczone do ogólnej liczby wyników" (16:07:56+0100)), changes in reporting over time and variation in testing regimes (e.g. in Germany from the 11th of November on Ärzteblatt (2020)).

Line list data used to inform the delay from symptom onset to test postive case report or death in the model based forecasts was sourced from (cite public linelist) with data available up to June (check exact date). Population data at the national and state level in Germany and Poland used in the model based forecasts was sourced from (source for population data).

Forecasts

Model based forecasts We used two models from the `EpiNow2` R package (version 1.3.3) as our baseline model based forecasts (Abbott, Hellewell, et al. 2020). These were chosen for their relative simplicity, attention to modelling the observation model of the forecast targets, and their grounding in simplistic epidemiological assumptions. The first of these models, which was used to forecast both test positive cases and deaths, used the renewal equation (Cori et al. 2013b) and an approximate Gaussian process (Riutort-Mayol et al. 2020) to estimate the effective reproduction number over time for latent infections and then convolved these infections to the target observation using data based delay distributions "Evaluating the Use of the Reproduction Number as an Epidemiological Tool, Using Spatio-Temporal Trends of the Covid-19 Outbreak in England | medRxiv" (n.d.). The second model, which was only used to forecast

deaths, assumed that deaths could be modelled using a scaling parameter, a convolution of test positive cases with a distribution that described the delay from case report to death, and a negative binomial observation model with a day of the week effect (Abbott, Hellewell, et al. 2020). Both models are described in detail in the supplementary information.

Each forecast target was fit independently for each model using Markov-chain Monte Carlo (MCMC) in stan (Stan Development Team 2020). A minimum of 4 chains were used with a warmup of 250 samples for the renewal equation based model and 1000 samples for the convolution model. 2000 samples total post warmup were used for the renewal equation model and 4000 samples of the convolution model. Different settings were chosen for each model to optimise compute time contigent on convergence. Convergence was assessed using the R hat diagnostic (Stan Development Team 2020). For the convolution model forecast the case forecast from the renewal equation model was used in place of observed cases beyond the forecast horizon using 1000 posterior samples.

Crowd forecast Crowd forecasts were created by ensembling forecasts submitted by individual participants. Participants were recruited mostly within the Centre of Mathematical Modeling of Infectious Diseases at the London School of Hygiene and Tropical Medicine, but participants were also invited personally or via social media to submit predictions.

Collection Participants were asked to make forecasts of Covid-19 cases and deaths over a four week ahead horizon using a web application (https://cmmid-lshtm.shinyapps.io/crowd-forecast/). The application was built using the shiny and golem R packages (Chang et al. 2021; Fay et al. 2021) and is available in the `crowdforecastr` R package (N. I. Bosse et al. 2020). Forecasts could be made until Tuesday, but forecasters were only allowed to use data available up until Monday. To make a forecast in the application participants could select a predictive distribution, with the default being log-normal, and adjust the median and the width of the uncertainty by either interacting with a figure showing their forecast or providing numerical values. The baseline shown was a repetition of the last known observation with constant uncertainty around it computed as the standard deviation of the last four observed log changes in forecasts. We required that participants submitted forecasts with uncertainty that increased over time. Our interface also allowed participants to view the observed data, and their forecasts, using a log scale and presented additional contextual COVID-19 data sourced from ourworldindata.org ("COVID-19 Data Explorer" n.d.). These data included notifications of both test positive COVID-19 cases and COVID-19 linked deaths, case fatality rates and the number of COVID-19 tests though the availability of the data evolved over the study period.

Processing Forecasts were downloaded, cleaned and processed every week for submission. If a forecaster had submitted multiple predictions for a single target, only the latest submission was kept. Some personal information (like the exact

time of the forecast) was removed. Information on the chosen distribution as well as the parameters for median and width were used to obtain a set of 23 quantiles from that distribution. Forecasts from all eligible forecasters were then aggregated using an unweighted quantile-wise mean (citation mean ensembles are good). In the beginning, inclusion was decided based on the authors' ad-hoc assessment of the validity of the forecast submission. Almost all forecasts were kept if they weren't clearly a result of someone experimenting with the app. From XX we based inclusion on the criterion that a forecaster submitted forecasts for at least two targets.

Forecast submission

Both computer generated forecasts and crowd preditions were submitted every Tuesday 3pm (using data up until Monday) for a one to four week ahead horizon. Forecasts were submitted in a quantile-based formats with 22 quantiles plus the median prediction.

All forecasts were processed in a Docker container that ran automated cron jobs to ensure a reproducible environment. All code and tools necessary to generate the forecasts and make a forecast submission are available in the covid.german.forecasts R package.

All forecasts are available here: https://github.com/epiforecasts/covid.german.forecasts

Statistical analysis

Forecasts were analysed by visual inspection as well formal model evaluation. Forecast submissions were visualised by forecast time horizon and compared to the ensemble of all forecasts from the German and Polish Forecast Hub. Formal model evaluation was based on the weighted interval score (wis) (Johannes Bracher et al. 2021), as well as empirical coverage of the 50% and 90% prediction intervals. The WIS was used to compute a relative skill value (smaller means better) that takes all possible pairwise comparisons between models into account and therefore provides a relative model ranking. Relative skill for all models was divided by the relative skill achieved by the baseline model to obtain a scaled relative skill value. All scores were calculated using the `scoringutils` package (N. Bosse 2020) in R. For case forecasts, all forecasts from October 12th 2020 until March 1st 2021 were taken into account. For deaths, we only score forecasts made after the 14th December, as no fully operational version of the convolution model was available before.

For the main analysis we focused on two week ahead predictions, as this is the horizon most commonly focused on (J. Bracher et al. 2021). Forecasts more than two weeks are often very unreliable as conditions change rapidly. Forecast scores for other horizons are given in the Appendix. Scores were aggregated by target type (deaths or cases) in the table, but plotted for every country separately to give a more detailed overview. In addition we also stratified the time series into three different categories for every forecast date. Depending on whether numbers

were monotonically rising or falling over the last two weeks prior to a given forecast date (i.e. whether the last three points formed a monotonically rising or falling line), the epidemic was categorised as either 'increasing,' 'decreasing' or 'unclear.' Differences of less than 5% relative to the week before were treated as zero in the classification, meaning they were interpreted as consistent with either classification.

Results

Forecast submission

Forecasts were submitted every Tuesday, 3pm. The model based forecasts used data up to the previous Sunday. Human forecasters were allowed to make forecasts on Tuesday, but were asked to use only information up to Monday. Before the 7th of December, only the renewal model and the crowd forecasts were submitted. Starting with the 7th of December, the convolution model was added. As the first submission suffered from a software bug, we excluded it from this analysis. March 1st was chosen as the last submission date, as we switched to submitting to the European Forecast Hub on the 8th of March (which involved changing the data source). From the XXth on we also submitted model based forecasts on a regional level. These forecasts were not further analysed as we could not produce corresponding crowd forecasts due to the large number of locations. Model based forecasts used the same approach throughout the forecast period with no changes to the methodology or setting. Interventions that applied at different points throughout the study period were therefore not explicitly modelled. Human forecasters were of course able to adapt their forecasts to current or likely future interventions.

A total number of 31 participants have submitted forecasts (although duplicates cannot be ruled out). The median number of forecasters was 6, the minimum 2 and the maximum 9 for a single forecast target. Participation rose steadily and peaked in February, before declining again towards the end of the study period. Motivating forecasters to contribute regularly proved challenging. The mean number of submissions from an individual forecaster was 4.7, but the median number was only one - most participants unfortunately dropped out after their first submission. Only two participants submitted a forecast every single week. To increase usability, the interface and its visual appearance was continuously tweaked and improved and additional information, e.g. from ourworldindata.org was added. The core functionality, however, remained unchanged.

Comparison of forecast performance

Summarised scores are given in Table 1, and a visualisation of two week ahead forecasts and scaled relative skill scores for the different models over time is shown in Figure 1. Model performance was quite varied across time and prediction targets. For cases, crowd forecasts were able to perfom en par with the ensemble of all forecasts from the Hub. The renewal model performed noticeably worse,

owing to a few predictions that were far away from the observed data. The renewal model had a general tendency to overpredict that was especially pronounced in situations where case numbers were growing on the date of the forecast. Figure 3 illustrates the relative contributions of sharpness, over- and underprediction to the WIS in different situations. All models showed poor calibration with respect to the empirical coverage of the 50% and 90% prediction intervals. Notably, only 55% of observations fell within the 90% prediction interval of the crowd forecast. It was, however, best in terms of the absolute error, suggesting that human forecasters were able to predict the general trend reasonably well, but failed to calibrate their own uncertainty and were overly confident in their predictions. The visualisation of forecasts in Figure 1 suggests that human forecasters may be slightly better than model based forecasts to predict changes in trend. In November, for example, they were able to predict a slowdown in the number of cases in Germany and Poland more accurately than the model based forecasts.

With regards to death forecasts, the ensemble clearly outperformed all other models and overall made unbiased and well calibrated predictions. The poorer relative performance of the crowd forecasts suggest that humans may be at a relative disadvantage when it comes to forecasting deaths. Future cases depend on a large degree on countless factors that are very hard to model (e.g. future policy interventions, adherence to rules, seasonal effects) but that can potentially be taken into account by humans. For deaths, the direct relationship to past case numbers is more pronounced and can to a certain exent be captured using epidemiological modelling. From an epidemiological point of view, deaths, for example, can simplistically be understood as a convolution of case numbers on some delay distribution multiplied by a case fatality rate. These relationships are hard for humans to untangle and quantify who can only rely on their intuition to make predictions. On deaths, the convolution model outperformed the Renewal model, implying that it was easier to model deaths as a convolution of case numbers, instead of estimating a separate death forecast based on a separate Rt value. It should be noted that the convolution model based its predictions on the forecast of the renewal model, which often struggled with case predictions. Its performance therefore needs to be interpreted accordingly. Possible future research could look into how a convolution model performs that is based on a more accurate case forecast.

Model	Target	Skill	WIS	Sharpness	Underpred.	Overpred.	Bias	Abs. error	50%-Cov.	90%-Cov.
Crowd forecast	case	0.80	16200	3660	5930	6600.0	-0.01	23300	0.36	0.55
Hub-ensemble	case	0.81	16500	5450	3290	7710.0	0.02	24300	0.43	0.69
Baseline	case	1.00	20200	4750	10000	5490.0	-0.06	28400	0.31	0.55
Renewal	case	1.26	25600	5420	5920	14200.0	0.17	34600	0.43	0.67
Hub-ensemble	death	0.62	296	125	91	80.2	0.05	486	0.58	0.92
Convolution	death	0.74	357	104	204	48.8	-0.10	565	0.33	0.79
Crowd forecast	death	0.77	368	107	102	160.0	0.14	576	0.38	0.75
Baseline	death	1.00	479	123	122	233.0	0.23	735	0.17	0.67
Renewal	death	1.10	524	155	133	236.0	-0.02	750	0.50	0.71

Table 1: Scores for 2 week ahead forecasts (cut to three significant digits and rounded). Skill is the scaled relative skill, a measure of relative performance with

respect to the baseline model (lower values are better). Sharpness, overprediction and underprediction together some up to the weighted interval score (WIS, lower values are better). 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.

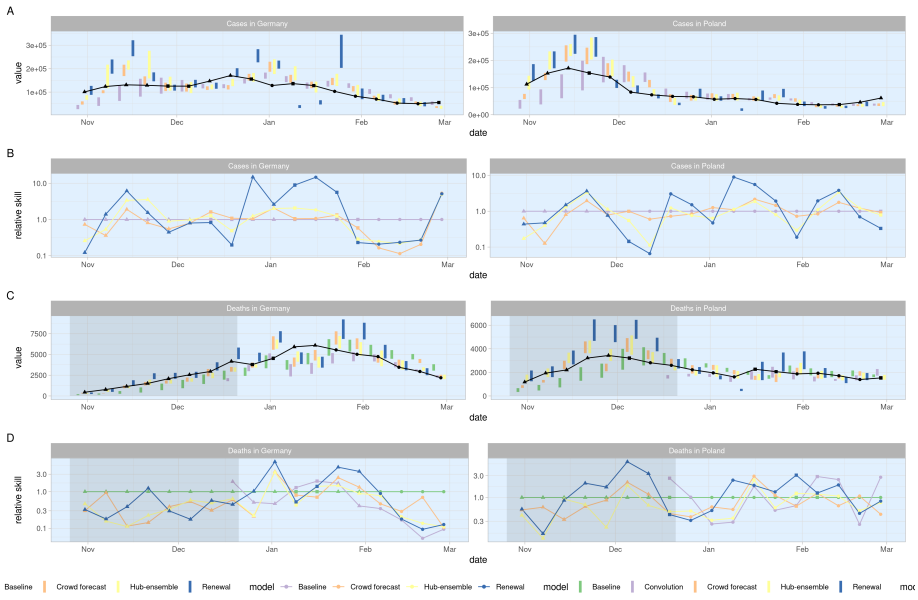


Figure 1. A, C: Visualisation of 2 week ahead forecasts against the true observed values. The shape indicates whether there has been a monotonic increase or decrease over the last two weeks leading up to a given data point, or an unclear trend. Forecasts that aren't scored (because there was no complete set of death forecasts available) are greyed out. B, D: Visualisation of corresponding scaled relative skill scores for the forecasts shown on the left. Scaled relative skill scores can be thought of as 'improvement over the baseline model' (see Methods for details). The shape indicates whether the trend was rising, falling or unclear at the date when the corresponding forecast was made (i.e. two weeks earlier)

Figure 2 shows the distribution of scaled relative skill scores achieved by each model. The renewal model had by far the largest variance in terms of its performance. Across different targets, countries and phases the distirbution of its scores tended to be bimodal, meaning that forecasts tend to be either really good, clearly outperforming the baseline and often most other models, but also frequently very far off. Among all models, the crowd forecasting model had the lowest variance in scores and performed most consistently.

As is illustrated in panel A in Figure 2, median forecasts from all models beat the baseline on the two week horizon displayed here (note that as shown in Table 1 mean forecasts from the renewal model performed worse than the baseline). Models outperformed the baseline more easily for death forecasts than for case forecasts, reinforcing the notion that deaths are easier to forecast than cases (E. Y. Cramer et al. 2021); J. Bracher et al. (2021)).

For the renewal and the convolution model, performance varied more strongly across countries than for the ensemble and crowd forecast model. Interestingly, the convolution model performed well on deaths in Germany, even though the model performed poorly. Conversely, it performed poorly on deaths in Poland even though the renewal had performed relatively well.

Interpretation of the different phases - Discuss with Sam - Renewal and Crowd forecast good when cases are rising (maybe: renewal model is good at modelling exponential growth and humans are good at adapting to turning points) - hard to identify a clear pattern

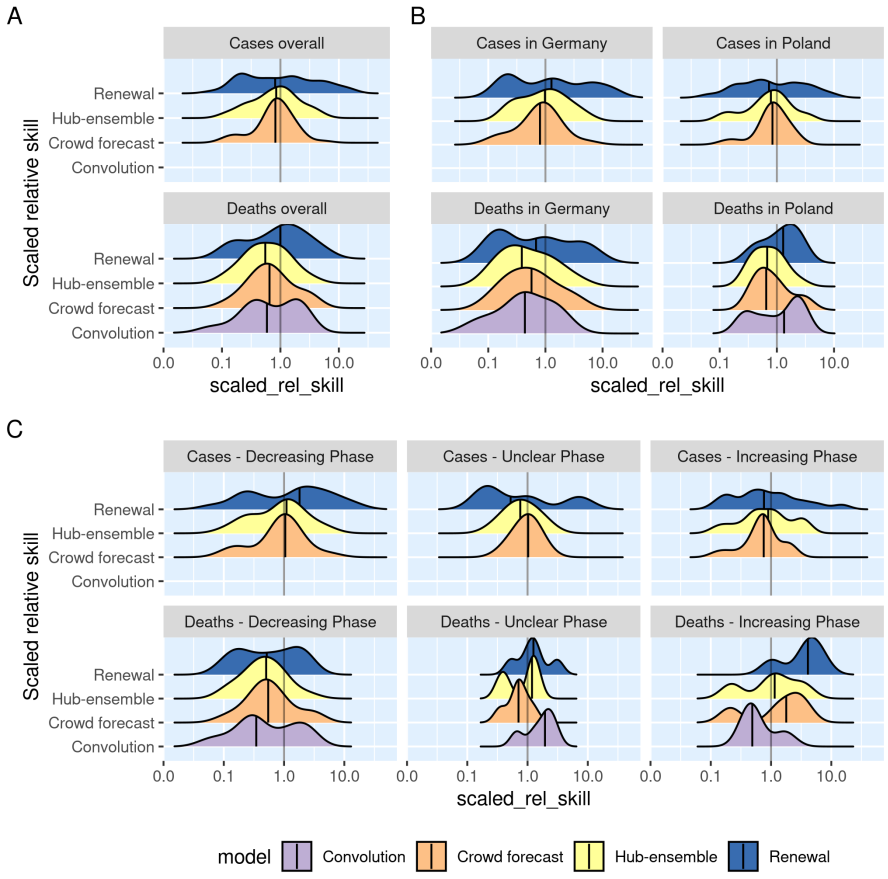


Figure 2. A: Overall distribution of the scaled relative skill scores (smaller is better) for the different models and forecast targets. The vertical black line at x = 1 represents the baseline model. B: Distribution of scaled relative skill scores separate by country. C: Distribution of scaled relative skill in different phases of the epidemic. Phases are classified according to whether the two weeks prior to the date when a forecast was made show a consistent trend.

All forecasts deteriorated with increasing forecast horizons, albeit at different rates. Figure 3 shows the distribution of the WIS for all models, locations and horizons. While mean and median performance was generally good for the renewal model one week ahead, it quickly deteriorated with increasing forecast horizon.

Figure 3. Distribution of weighted interval scores achieved by the models at different horizons. Mean performance (black circles) was generally worse than median performance (black squares), implying that the distribution is skewed and suffers from outliers where models make predictions far away from the true observed values. **Maybe make an 'overall' panel as well**

Relative contributions to the weighted interval scores changed depending on the phase of the epidemic. Generally, models tended to underpredict when cases or deaths where increasing, and overprdict when cases and deaths were decreasing. Especially when case numbers are rising, there is a large danger of overshooting and missing a change in trend. This danger is asymmetric when looking at numbers on a linear scale instead of a log scale, as numbers are bound by zero, but can grow very large. To a certain extent, underpredicting may be interpretable

as 'hedging against' or incorporating the fact that a sudden downturn may be possible. Given that underpredictions made up a large part of penalties incurred during increasing phases this implies that either humans were not well prepared to forecast exponential growth in cases or systematically overestimated the probably of a sudden downturn. For deaths it was even more striking that all models consistently overpredicted deaths, maybe missing a change in the observed case fatality rate due to changes in testing. It is interesting to see that the pattern of the relative share of the WIS components for human forecasters most closely resembles the pattern of the baseline model.

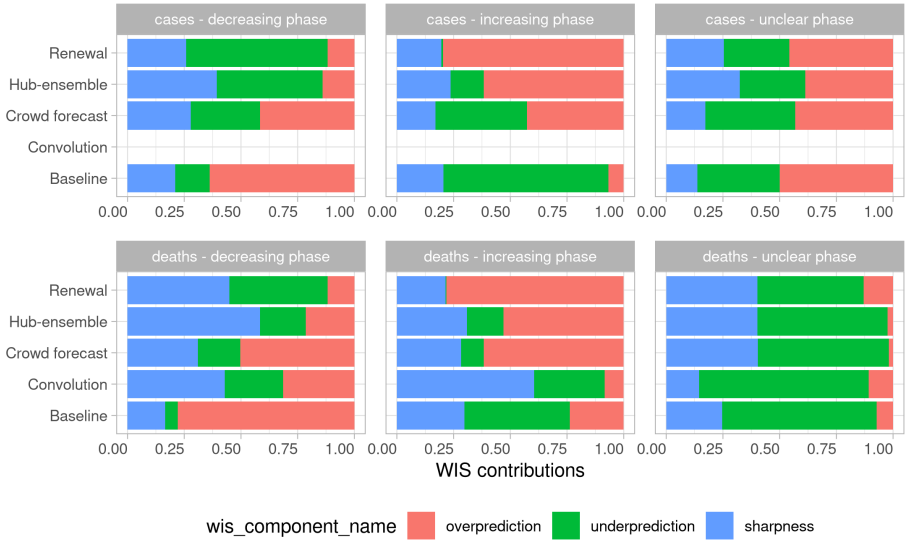


Figure 3. Relative contributions of sharpness, over- and underprediction to the overall weighted interval score achieved by a model in different phases of the epidemic. Note that the uncertainty of the baseline model depends on the variation of observed differences in the past and is therefore naturally hihger in an unclear phase.

Discussion

Summary

Any model forecasts are informed by mechanistic model assumptions as well as the researcher's subjective opinion that shapes they way a model is tweaked and tuned. Expert judgement alone (in the form of aggregated crowd forecasts) can make a valuable contribution to Covid-19 forecasting. Participants, however, were difficult to recruit. In addition, forecasting a lot of different targets is strenuous for individual forecasters, limiting the number of possible forecast targets. We therefore could not produce crowd forecasts at the state level in either Poland or Germany due to a lack of researcher time and our ability to

reach out to potential forecasters. The fact that most participants were from the UK and had no connection to Germany or Poland made motivating them even more difficult. It also severely limited the amount of domain knowledge and expertise among our forecasters. This suggests that the performance of the crowd forecast ensemble is likely closer to the lower than to the upper bound of what could reasonably be expected of an expert forecast.

With regards to cases our crowd forecasts clearly outperformed the untuned renewal model and performed on average as well as or better than a large ensemble of model derived, and expert tuned, forecasts. This suggests (*does it?*) that expert judgement may often play a large and important role even in forecast models that on the surface derive their predictive ability from epidemiological theory. Humans are relatively good at foreseeing trends in case numbers which are influenced by many factors that are hard to model (such as likely future interventions, adherence to policy interventions or unknown seasonal effects). However, humans tended to be even more overconfident in their predictions than model based forecasts.

Compared to human forecasters, mechanistic models are at a relative advantage when predicting deaths from Covid-19. The ensemble forecast clearly outperformed crowd forecasts and provided unbiased and well calibrated predictions. Notably, a simple model that assumed that cases were a convolution of deaths performed well compared to other approaches. The convolution model outperformed expert opinion at short time horizons and did relatively well at longer time horizons. This was the case even though the convolution model was based on case forecasts from the renewal model which often performed poorly.

The renewal model which used daily data was good at predicting exponential growth or decline and adapted well to changes in trend. It performed well at short horizons, but suffered at times from severe overprediction and a general tendency to deteriorate over longer forecast horizons. Whether or not that is acceptable depends on the use case of those who are consuming the forecasts. While the German and Polish Hub asked for one to four week ahead predictions it is not clear what forecast horizon the predictions should be optimised for (and evaluated against) without further context.

Strengths and Weaknesses

When writing this layer strengths and limitations together. Start with a strength and then for every limitation counter with a strength

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

evaluations.

While the methodology did not change for the Renewal model, the Convolution model and the Baseline model, this continuity is not given for the crowd forecasts and the hub ensemble model. Comparability of crowd forecasts at different time points is hampered by the low number of participants we were able to recruit initially and the fact that participants kept joining or dropping out. Similarly, the composition of the Hub ensemble changed over time as did many of the individual models contributing forecasts to the Forecast Hub. To mitigate this, a wide range of potential confounding factors, like different time periods, were considered to ensure the robustness of the obtained results.

Human forecasters performed surprisingly well in spite of the low number of contributors and the limited knowledge of the situation in Germany and Poland among participants. Due to the small number of participants we were not able to easily compare the performance of experts versus non-experts. It is, however, plausible to assume that an ensemble of a larger number of experts who have genuine knowledge not only in epidemiology but also in the country for which they make a forecast, would show improved performance. In that sense our crowd forecasts established a baseline that is likely at the lower end of what can reasonably achieved with human expert forecasting. The fact that the interface demanded some understanding of distributions and time series made it hard to recruit participants who didn't already have a background in either some quantitative field or epidemiology. On the other hand, the setup allowed us to obtain a full predictive distribution instead of only a limited set of quantiles. This made it possible to compare expert forecasts directly against computer generated forecasts using the same evaluation tools. While forecasts performed well for a limited set of targets the setup is not easily scalable to a large set of prediction targets. Using R shiny also came with some limitations in terms of usability. On the other hand, setting the platform up as a public R package means that it can easily be adapted and re-used for future forecasting projects.

Strengths and weaknesses in the context of the literature

Same structure as for the previous section.

Other efforts have attempted to compare forecasts of Covid-19 submitted by different research group. A comparison of model performance in the US (E. Y. Cramer et al. 2021) had a much larger data set of forecast targets and models. On the one hand, this allowed for more robust statistical inference. On the other hand, a large number of models and targets makes it more difficult to draw conclusions that go beyond a ranking of models. Models in the US Forecast Hub essentially had to be treated as a black box, as not all details were known (or collecting them was infeasible). In addition no human forecasts directly entered the models analysed in E. Y. Cramer et al. (2021). Focusing on a small number of known models and forecast targets allowed us to obtain a deeper understanding of how models and human forecasts performed. J. Bracher et al. (2021) published an evaluation of all forecasts submitted to the German and

Polish Forecast Hub. Their study was pre-registered, ensuring full transparency of the results obtained. They also included different interventions in their analysis and the effect they may have had on scores. This was not feasible for us to do, as especially a list of Polish interventions was not readily available to us. In addition it is not entirely clear what constitutes an 'intervention' and there are many researcher degrees of freedom involved. Instead, we decided to categorise the time series in 'rising,' 'falling' and 'unclear' and therefore implicitly looked at whether models were able to foresee future interventions or other factors that may lead to a change in trend.

Other crowd forecasting projects like the Delphi project [no citation found], the expert elicitation efforts led by Thomas McAndrew (McAndrew et al. 2021), Metaculus ("A Preliminary Look at Metaculus and Expert Forecasts" n.d.) or Good Judgement Open (Tetlock et al. 2014) often had a far greater number of participants. However, their forecasts are not directly comparable to model based forecasts. The pool of participants is also much different from the modellers who usually submit predictions to Covid-19 Forecast Hubs. In our case many of our forecasters came from the same modelling group that also submitted model based forecasts, allowing us to more clearly disentangle the contributions of human judgement and model derived insights.

Future work

The work presented here can and should be expanded in various ways. For the purpose of this paper, crowd forecasts were treated as a a single model, where in reality they are are an ensemble of very different individual opinions. There are various ways in which these opinions could be combined into a single forecast that we intend to explore in the future, for example weighted ensembles that give forecasters more weight who performed well in the past. Investigating *why* successful forecasters predicted numbers to rise or fall is also likely to yield insights that can be useful for policy makers. Another promising research project is a forecast that combines human opinion with epidemiological modelling. In the UK we are currently asking humans to predict Rt instead of cases and deaths directly. From the projected Rt trajectory we simulate cases as well as deaths which means that humans can focus on the overall trend, while models deal with the mechanistic details. We hope to improve human death forecasts substantially using this method.

Conclusions

- Crowd forecasts outperform models with simplistic epidemiology derived assumptions at longer time horizons.
- At shorter time horizons performance is more comparable, especially when forecasting deaths when a model that simplistically assumes that deaths are scaled convolution of cases performs relatively well.

Todo

Decisions and things to think about

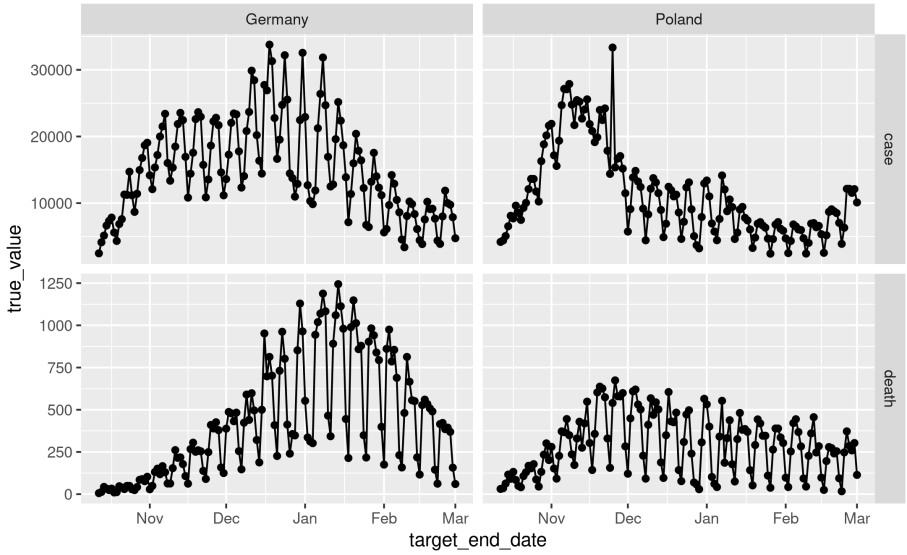
- which dates, if any, do we want to exclude? Daily forecasts don't look too irregular, except for the spike of cases in Poland
- Can we recalculate the ensemble excluding our models?
- Figure 2: maybe we don't want to stratify according to location? –> idea is that we don't do it for the table. We also don't really talk about particular interventions in Germany and Poland, so maybe it would make sense to stick with 'cases' and 'deaths' as the two distinct categories and avoid introducing more noise by stratifiying it according to country
- Same reasoning for Figure 3?
- For Figure 1 we might want to rearrange the plots such that the forecast is always above the score. I.e. have it case case score score deaths deaths score score
- maybe move Figure 3 to the Appendix?
- in general be clearer: are we a crowd forecast or an expert forecast?
- maybe look more into the composition of forecasters - e.g. we had some Germans, but no one from Poland

Actual todos

- I didn't really find a citation for the Delphi Crowdcast project. Could site this website? https://delphi.cmu.edu/crowdcast/ but that is just the url of the app...

Supplementary information

Daily forecasts



Visualisation of daily report data. Issue with this is that it isn't data os of then, but as of now. ## Forecast models

Effective Reproduction number model

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

$$I_t = I_0 \exp(rt) \tag{1}$$
$$I_0 \sim \mathcal{LN}(\log I_{obs}, 0.2) \tag{2}$$
$$r \sim \mathcal{LN}(r_{obs}, 0.2) \tag{3}$$
$$\tag{4}$$

Where I_{obs} and $r_{obs} - 12 weeks of data - Prior log-normal with a mean of 1.1 and a standard deviation of 0.2 - Days with missing data or 0 notifications adjusted to the 7 day moving average if the 7 day moving average of notifications was greater than 50 per day. - Population adjustment (cite epidemia) - Assumed static normally distributed reporting fraction with a mean of 0.25 and a standard deviation of 0.05 for test positive cases and a mean of 0.005 with a standard deviation of 0.0025 for COVID-19 linked deaths. - Rt fixed from the forecast horizon. - 4 chains, 250 warmup samples per chain, and 2000 samples overall post warmup.

We estimated the instantaneous reproduction number (R_t) using the `EpiNow2` R package (version 1.2.1) (Abbott, Hellewell, et al. 2020) on the last 12 weeks of available data

The instantaneous reproduction number represents the number of secondary cases arising from an individual showing symptoms at a particular time, assuming that conditions remain identical after that time, and is therefore a measure of the instantaneous transmissibility (in contrast to the case reproduction number - see Fraser (2007) (Fraser 2007) for a full discussion). `EpiNow2` implements a Bayesian latent variable approach using the probabilistic programming language Stan (Stan Development Team 2020), which works as follows. The initial number of infections were estimated as a free parameter with a prior based on the initial number of cases, or deaths, respectively. The initial, unobserved, growth rate was estimated from the first 7 days of reported data. This was used as a prior (normal with standard deviation 0.2) to estimate latent infections prior to the first reported case using a log linear model. For each subsequent time step, previous imputed infections (I_{t-1}) were summed, weighted by an uncertain generation time probability mass function (w), and combined with an estimate of R_t to give the incidence at time t (I_t) (Abbott, Hellewell, et al. 2020; Cori et al. 2013a; Thompson et al. 2019). We used a log normal prior for the reproduction number (R_0) with mean 1 and standard deviation 0.2 reflecting our current belief that R_t is likely to be centred around 1 in most of the world, with public health interventions and individual behaviour combining to prevent it from growing significantly larger for sustained periods.

The infection trajectories were then mapped to mean reported case counts (D_t) by convolving over an uncertain incubation period and report delay distribution (convolved into ξ). Observed reported case counts (C_t) were then assumed to be generated from a negative binomial observation model with overdispersion ϕ (using 1 over the square root of a half normal prior with mean 1) and mean D_t, multiplied by a day of the week effect with an independent parameter for each day of the week ($\omega_{(t \mod 7)}$). Temporal variation was controlled using an approximate Gaussian process (Riutort-Mayol et al. 2020) with a squared exponential kernel (GP). In mathematical notation,

This package implements a Bayesian latent variable approach using the probabilistic programming language Stan (27). To initialise the model, infections were imputed prior to the first observed case using a log linear model with priors based on the first week of observed cases. This means that the initial observations both inform the initial parameters and are then also fit, which makes the initial Rt estimates less reliable than later estimates. This was a pragmatic choice to allow the model to be identifiable when only estimating part of the observed epidemic. We explored other parameterisations, but these suffered from poor model identification. For each subsequent time step with observed cases, new infections were imputed using the sum of previous modelled infections weighted by the generation time probability mass function, and combined with an estimate of Rt, to give the prevalence at time t (12). The generation time was assumed to

follow a gamma distribution that was fixed over time but varied between samples, with priors drawn from the literature for the mean and standard deviation (28).

$$I_{t_{unobserved}} = I_0 \exp\left(rt_{unobserved}\right) \tag{5}$$
$$I_0 \sim \mathcal{LN}(\log I_{observed}, 0.2) \tag{6}$$
$$r \sim \mathcal{LN}(r_{observed}, 0.2) \tag{7}$$
$$\tag{8}$$

$$\log R_t = \log R_{t-1} + \mathrm{GP}_t \tag{9}$$
$$I_t = R_t \sum_{\tau}^{15} w(\tau|\mu_w, \sigma_w) I_{t-\tau} \tag{10}$$
$$O_t = \sum_{\tau}^{15} \xi_O(\tau|\mu_{\xi_O}, \sigma_{\xi_O}) I_{t-\tau} \tag{11}$$
$$D_t = \alpha \sum_{\tau}^{15} \xi_D(\tau|\mu_{\xi_D}, \sigma_{\xi_D}) O_{t-\tau} \tag{12}$$
$$C_t \sim \mathrm{NB}\left(\omega_{(t \mod 7)} D_t, \phi\right) \tag{13}$$

Where,

$$R_0 \sim \mathcal{LN}(0.079, 0.18) \tag{14}$$
$$w \sim \mathcal{G}(\mu_w, \sigma_w) \tag{15}$$
$$\xi_O \sim \mathcal{LN}(\mu_{\xi_O}, \sigma_{\xi_O}) \tag{16}$$
$$\xi_D \sim \mathcal{LN}(\mu_{\xi_D}, \sigma_{\xi_D}) \tag{17}$$
$$\tag{18}$$

with the following priors,

$$\mu_w \sim \mathcal{N}(3.6, 0.7) \tag{19}$$
$$\sigma_w \sim \mathcal{N}(3.1, 0.8) \tag{20}$$
$$\mu_{\xi_O} \sim \mathcal{N}(1.62, 0.064) \tag{21}$$
$$\sigma_{\xi_O} \sim \mathcal{N}(0.418, 0.069) \tag{22}$$
$$\mu_{\xi_D} \sim \mathcal{N}(0.614, 0.066) \tag{23}$$
$$\sigma_{\xi_D} \sim \mathcal{N}(1.51, 0.048) \tag{24}$$
$$\alpha \sim \mathcal{N}(0.25, 0.05) \tag{25}$$
$$\frac{\omega}{7} \sim \text{Dirichlet}(1, 1, 1, 1, 1, 1, 1) \tag{26}$$
$$\phi \sim \frac{1}{\sqrt{\mathcal{N}(0, 1)}} \tag{27}$$

When forecasting deaths the following alternative priors were used,

$$\mu_{\xi_D} \sim \mathcal{N}(2.29, 0.076) \tag{28}$$
$$\sigma_{\xi_D} \sim \mathcal{N}(0.76, 0.055) \tag{29}$$
$$\alpha \sim \mathcal{N}(0.005, 0.0025) \tag{30}$$
$$\tag{31}$$

α, μ, σ, and ϕ were truncated to be greater than 0 and with ξ, and w normalised to sum to 1. GP_t is an approximate Hilbert space gaussian process as defined in (Riutort-Mayol et al. 2020) using a Matern 3/2 kernel using a boundary factor of 1.5 and 17 basis functions (20% of the number of days used in fitting). The lengthscale of the Gaussian process was given a log-normal prior with a mean of 21 days, and a standard deviation of 7 days truncated to be greater than 3 days and less than 60 days. The magnitude of the Gaussian process was assumed be normally distributed centred at 0 with a standard deviation of 0.1. The prior for the generation time was sourced from (Ganyani et al. 2020) but refit using a log-normal incubation period with a mean of 5.2 days (SD 1.1) and SD of 1.52 days (SD 1.1) with this incubation period also being used as a prior (Lauer et al. 2020) for ξ_O. This resulted in a gamma distributed generation time with mean 3.6 days (standard deviation (SD) 0.7), and SD of 3.1 days (SD 0.8) for all estimates. We estimated the delay between symyptom onset and case report or death required to convolve latent infections to observations by fitting an integer adjusted log-normal distribution to 10 subsampled bootstraps of a public linelist for cases in Germany from April 2020 to June 2020 with each bootstrap using 1% or 1769 samples of the available data (Xu et al., n.d.; Abbott, Sherratt, et al. 2020) and combining the posteriors for the mean and standard deviation of the log-normal distribution (Abbott, Hellewell, et al. 2020; DOI n.d.; "Evaluating the Use of the Reproduction Number as an Epidemiological Tool,

Using Spatio-Temporal Trends of the Covid-19 Outbreak in England | medRxiv" n.d.; Stan Development Team 2020). This resulted in a delay distribution from symptom onset to case report with a mean of XX and a standard deviation of XX and a delay distribution from sypmtom onset to death with a mean of XX and a standard deviation of XX.

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

$$I_t = (N - I^c_{t-1}) \left(1 - \exp\left(\frac{-I'_t}{N - I^c_T}\right)\right), \tag{32}$$

where $I^c_t = \sum_{s<t} I_s$ are cumulative infections by $t-1$ and I'_t are the unadjusted infections defined above. This adjustment is based on that implemented in the `epidemia` R package (cite epidemia, cite (**bhatt202?**)).

Each forecast target was fit independently using using Markov-chain Monte Carlo (MCMC) in stan (Stan Development Team 2020). A minimum of 4 chains were used with a warmup of 250 each and 2000 samples total post warmup. Convergence was assessed using the R hat diagnostic (Stan Development Team 2020).

We used an estimate of the generation time sourced from .

Convolution model

- Summarise key choices
- data used
-

$$D_t \sim \text{NB}\left(\omega_{(t \mod 7)} \alpha \sum_{\tau=0}^{30} \xi(\tau|\mu,\sigma) C_{t-\tau}, \phi\right) \tag{33}$$

Where,

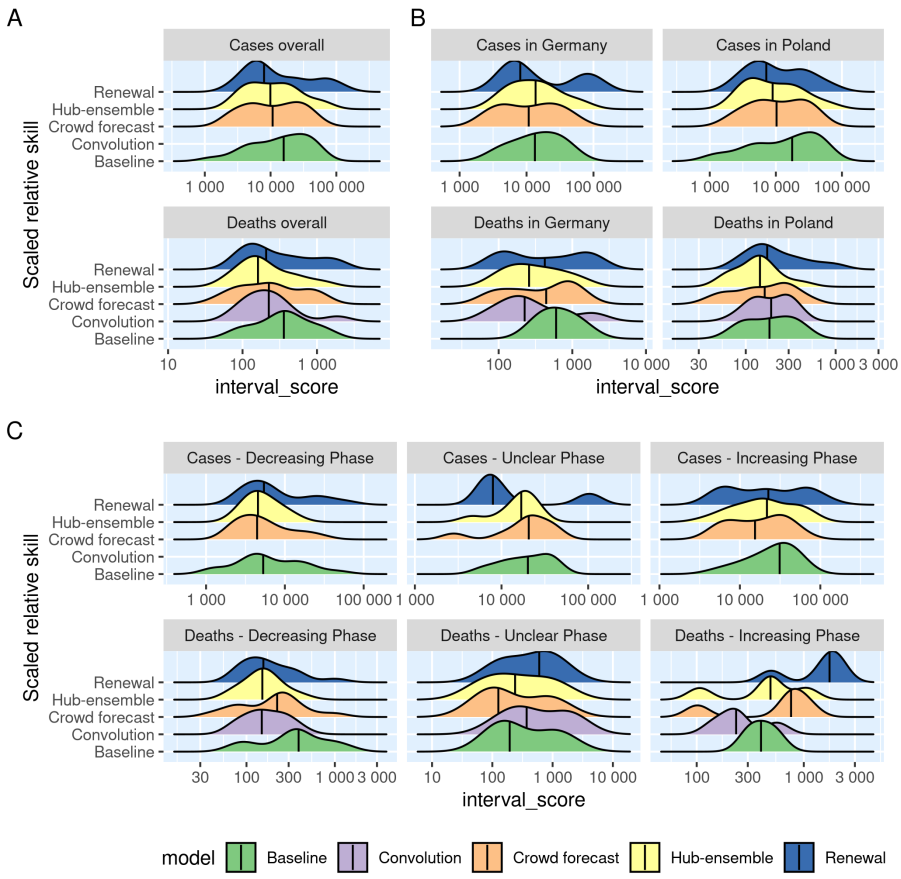
$$\frac{\omega}{7} \sim \text{Dirichlet}(1,1,1,1,1,1,1) \tag{34}$$
$$\alpha \sim \mathcal{N}(0.01, 0.02) \tag{35}$$
$$\xi \sim \mathcal{LN}(\mu, \sigma) \tag{36}$$
$$\mu \sim \mathcal{N}(2.5, 0.5) \tag{37}$$
$$\sigma \sim \mathcal{N}(0.47, 0.2) \tag{38}$$
$$\phi \sim \frac{1}{\sqrt{\mathcal{N}(0,1)}} \tag{39}$$

with α, μ, σ, and ϕ truncated to be greater than 0 and with ξ normalised such that $\sum_{\tau=0}^{30} \xi_(\tau|\mu,\sigma) = 1$. Only the last 3 weeks of data were included in the likelihood though all 12 weeks of data was used during fitting.

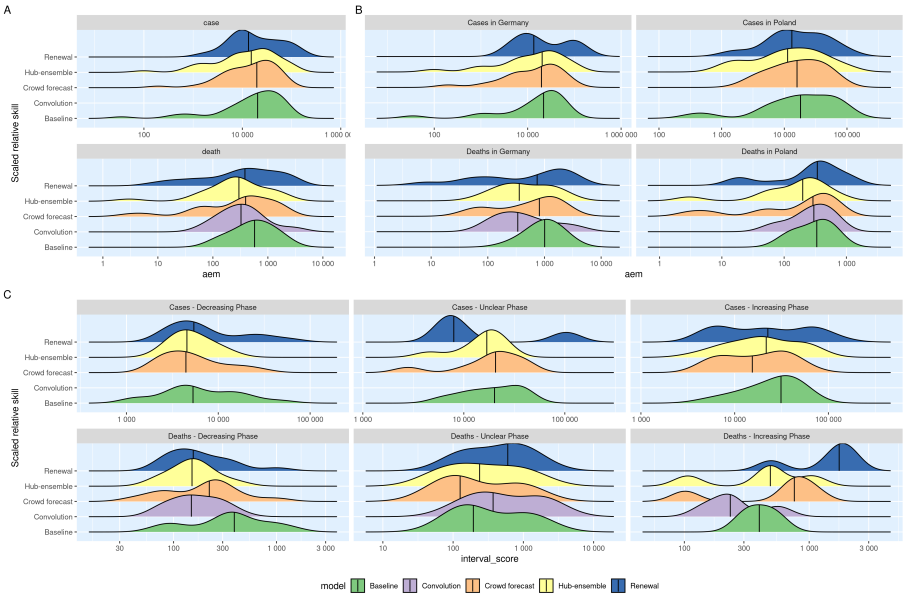
4 chains with 1000 warmup samples and 4000 posterior samples. 1000 posterior samples of the case forecast were then randomly matched with the posterior samples from the convolution model with the model being rerun for each sample to provide a forecast of future deaths.

Distribution of WIS

–> maybe kick these



This plot has wis instead of scaled relative skill



same plot with absolute error

References

"A Preliminary Look at Metaculus and Expert Forecasts." n.d. Accessed May 30, 2021. https://www.metaculus.com/news/2020/06/02/LRT/.

Abbott, Sam, Joel Hellewell, Joe Hickson, James Munday, Katelyn Gostic, Peter Ellis, Katharine Sherratt, et al. 2020. "EpiNow2: Estimate Real-Time Case Counts and Time-Varying Epidemiological Parameters." - - (-): –. https://doi.org/10.5281/zenodo.3957489.

Abbott, Sam, Katharine Sherratt, Jonnie Bevan, Hamish Gibbs, Joel Hellewell, James Munday, Patrick Barks, Paul Campbell, Flavio Finger, and Sebastian Funk. 2020. "Covidregionaldata: Subnational Data for the Covid-19 Outbreak." - - (-): –. https://doi.org/10.5281/zenodo.3957539.

Ärzteblatt, Deutscher Ärzteverlag GmbH, Redaktion Deutsches. 2020. "SARS-CoV-2-Diagnostik: RKI passt Testempfehlungen an." Deutsches Ärzteblatt. November 3, 2020. https://www.aerzteblatt.de/nachrichten/118001/SARS-CoV-2-Diagnostik-RKI-passt-Testempfehlungen-an.

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

Bosse, Nikos. 2020. *Scoringutils: A Collection of Proper Scoring Rules and*

Metrics to Assess Predictions. https://github.com/epiforecasts/scoringutils.

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

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

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

Bracher, Johannes, Evan L. Ray, Tilmann Gneiting, and Nicholas G. Reich. 2021. "Evaluating Epidemic Forecasts in an Interval Format." *PLoS Computational Biology* 17 (2): e1008618. https://doi.org/10.1371/journal.pcbi.1008618.

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

Cori, Anne, Neil M. Ferguson, Christophe Fraser, and Simon Cauchemez. 2013a. "A New Framework and Software to Estimate Time-Varying Reproduction Numbers During Epidemics." *American Journal of Epidemiology* 178 (9): 1505–12. https://doi.org/10.1093/aje/kwt133.

———. 2013b. "A New Framework and Software to Estimate Time-Varying Reproduction Numbers During Epidemics." *American Journal of Epidemiology* 178 (9): 1505–12. https://doi.org/10.1093/aje/kwt133.

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

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

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

DOI, Authors Affiliations Published Not published yet. n.d. "Covid-19: Temporal Variation in Transmission During the COVID-19 Outbreak." Covid-19. Accessed May 30, 2021. https://epiforecasts.io/covid/.

"Download Historical Data (to 14 December 2020) on the Daily Number of New Reported COVID-19 Cases and Deaths Worldwide." 2020. European Centre for Disease Prevention and Control. December 14, 2020. https:

//www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide.

"Evaluating the Use of the Reproduction Number as an Epidemiological Tool, Using Spatio-Temporal Trends of the Covid-19 Outbreak in England | medRxiv." n.d. Accessed May 30, 2021. https://www.medrxiv.org/content/10.1101/2020.10.18.20214585v1.

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

Ferguson, N., D. Laydon, G. Nedjati Gilani, N. Imai, K. Ainslie, M. Baguelin, S. Bhatia, et al. 2020. "Report 9: Impact of Non-Pharmaceutical Interventions (NPIs) to Reduce Covid19 Mortality and Healthcare Demand." Report. *20*. https://doi.org/10.25561/77482.

Fraser, Christophe. 2007. "Estimating Individual and Household Reproduction Numbers in an Emerging Epidemic." *PloS One* 2 (8): e758.

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

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

Ganyani, Tapiwa, Cecile Kremer, Dongxuan Chen, Andrea Torneri, Christel Faes, Jacco Wallinga, and Niel Hens. 2020. "Estimating the Generation Interval for Coronavirus Disease (COVID-19) Based on Symptom Onset Data, March 2020." *Eurosurveillance* 25 (17).

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

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

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

Lauer, Stephen A, Kyra H Grantz, Qifang Bi, Forrest K Jones, Qulu Zheng, Hannah R Meredith, Andrew S Azman, Nicholas G Reich, and Justin Lessler. 2020. "The Incubation Period of Coronavirus Disease 2019 (COVID-19) from Publicly Reported Confirmed Cases: Estimation and Application." *Annals of Internal Medicine* 172 (9): 577–82.

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

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

Riutort-Mayol, Gabriel, Paul-Christian Bürkner, Michael R. Andersen, Arno Solin, and Aki Vehtari. 2020. "Practical Hilbert Space Approximate Bayesian Gaussian Processes for Probabilistic Programming." https://arxiv.org/abs/2004.11408.

"RKI - Coronavirus SARS-CoV-2 - Aktueller Lage-/Situationsbericht Des RKI Zu COVID-19." n.d. Accessed May 30, 2021. https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Situationsberichte/Gesamt.html.

"Rozbieżności w statystykach koronawirusa. 22 tys. przypadków będą doliczone do ogólnej liczby wyników." 16:07:56+0100. 16:07:56+0100. https://forsal.pl/lifestyle/zdrowie/artykuly/8017628,rozbieznosci-w-statystykach-koronawirusa-22-tys-przypadkow-beda-doliczone-do-ogolnej-liczby-wynikow.html.

Stan Development Team. 2020. "RStan: The r Interface to Stan." http://mc-stan.org/.

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

Thompson, R. N., J. E. Stockwin, R. D. van Gaalen, J. A. Polonsky, Z. N. Kamvar, P. A. Demarsh, E. Dahlqwis, et al. 2019. "Impro@webpagekraemer2020epidemiological, Author = "Xu, Bo and Gutierrez, Bernardo and Hill, Sarah and Scarpino, Samuel and Loskill, Alyssa and Wu, Jessie and Sewalk, Kara and Mekaru, Sumiko and Zarebski, Alexander and Pybus, Oliver and Pigott, David and Kraemer, Moritz", Title = "Epidemiological Data from the nCoV-2019 Outbreak: Early Descriptions from Publicly Available Data", Url = http://virological.org/t/epidemiological-data-from-the-ncov-2019-outbreak-early-descriptions-from-publicly-available-data/337 Ved Inference of Time-Varying Reproduction Num-

bers During Infectious Disease Outbreaks." *Epidemics* 29: 100356. https://doi.org/https://doi.org/10.1016/j.epidem.2019.100356.

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

Xu, Bo, Bernardo Gutierrez, Sarah Hill, Samuel Scarpino, Alyssa Loskill, Jessie Wu, Kara Sewalk, et al. n.d. "Epidemiological Data from the nCoV-2019 Outbreak: Early Descriptions from Publicly Available Data." http://virological.org/t/epidemiological-data-from-the-ncov-2019-outbreak-early-descriptions-from-publicly-available-data/337.

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