Supplementary information

Weighted interval score

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

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

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

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

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

Renewal equation model

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

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

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

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

Where,

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

This model used the following priors for cases,

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

and updated the reporting process as follows when forecasting deaths,

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

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

The prior for the generation time was sourced from [1] 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 [2] for ξ_O. This resulted in a gamma-distributed generation time with mean 3.6 days (standard deviation (SD) 0.7), and SD of 3.1 days (SD 0.8) for all estimates. We estimated the delay between symptom onset and case report or death required to convolve latent infections to observations by fitting an integer adjusted log-normal distribution to 10 subsampled bootstraps of a public linelist for cases in Germany from April 2020 to June 2020 with each bootstrap using 1% or 1769 samples of the available data [3, 4] and combining the posteriors for the mean and standard deviation of the log-normal distribution [5, 6, 7, 8].

GP_t is an approximate Hilbert space Gaussian process as defined in [9] using a Matern 3/2 kernel using a boundary factor of 1.5 and 17 basis functions (20% of the number of days used in fitting). The length scale of the Gaussian process was given a log-normal prior with a mean of 21 days, and a standard deviation of 7 days truncated to be greater than 3 days and less than 60 days. The magnitude of the Gaussian process was assumed to be normally distributed centred at 0 with a standard deviation of 0.1. From the forecast time horizon (T) and onwards the last value of the Gaussian process was used (hence R_t was assumed to be fixed) and latent infections were adjusted to account for the proportion of the population that was susceptible to infection as follows,

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

where $I_t^c = \sum_{s<t} I_s$ are cumulative infections by $t-1$ and I'_t are the unadjusted infections defined above. This adjustment is based on that implemented in the `epidemia` R package [10, 11].

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

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

with the following priors,

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

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

Model fitting

Both models were implemented using the `EpiNow2` R package (version 1.3.3) [5]. Each forecast target was fitted independently for each model using Markov-chain Monte Carlo (MCMC) in stan [8]. A minimum of 4 chains were used with a warmup of 250 samples for the renewal equation-based model and 1000 samples for the convolution model. 2000 samples total post warmup were used for the renewal equation model and 4000 samples for the convolution model. Different settings were chosen for each model to optimise compute time contingent on convergence. Convergence was assessed using the R hat diagnostic [8]. For the convolution model forecast the case forecast from the renewal equation model was used in place of observed cases beyond the forecast horizon using 1000 posterior samples. 12 weeks of data was used for both models though only 3 weeks of data were included in the likelihood for the convolution model.

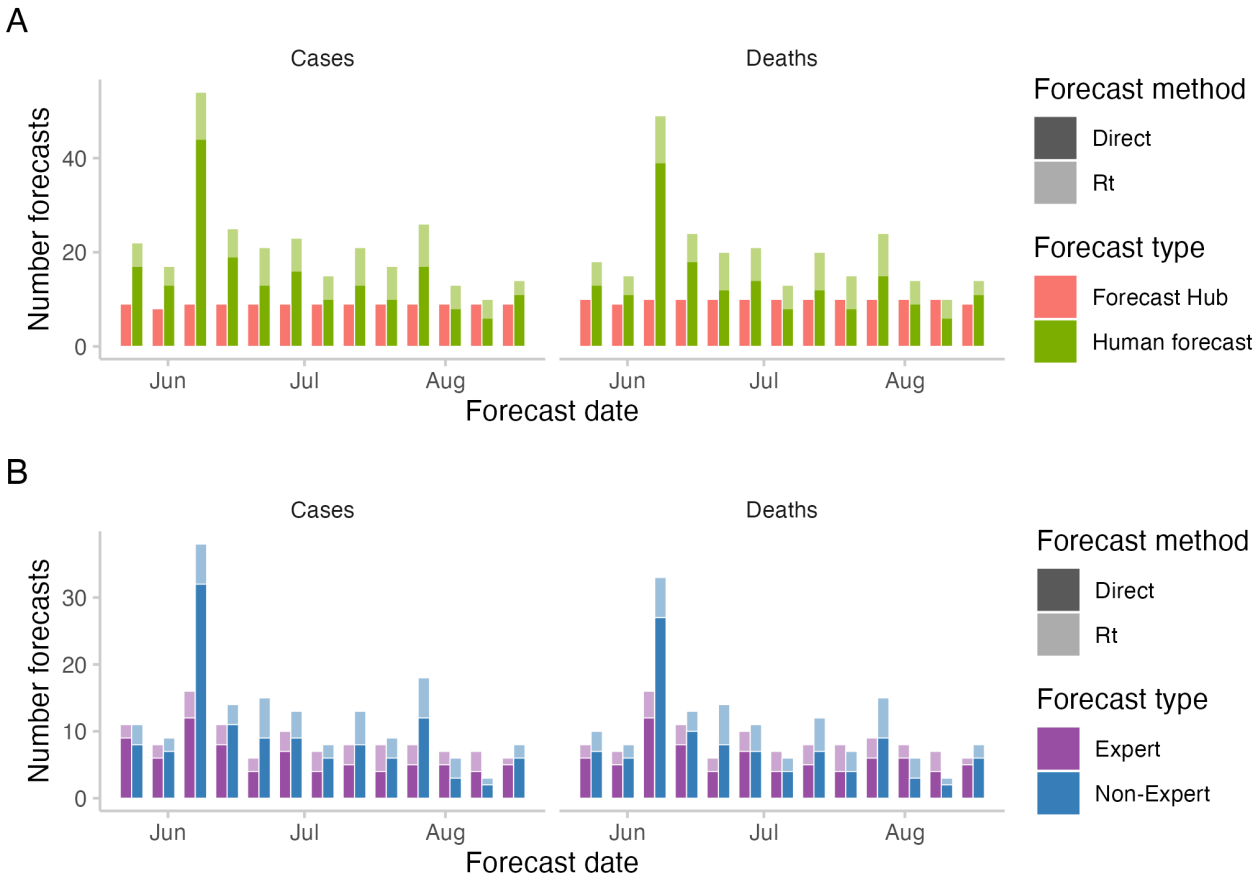

Figure SI.1. Number of forecasts across the study period. A: number of forecasts included in the Hub ensemble and the combined crowd ensemble. B: number of forecasts by "experts" and "non-experts". Expert status was determined based on the participant's answer to the question whether they "worked in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

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

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

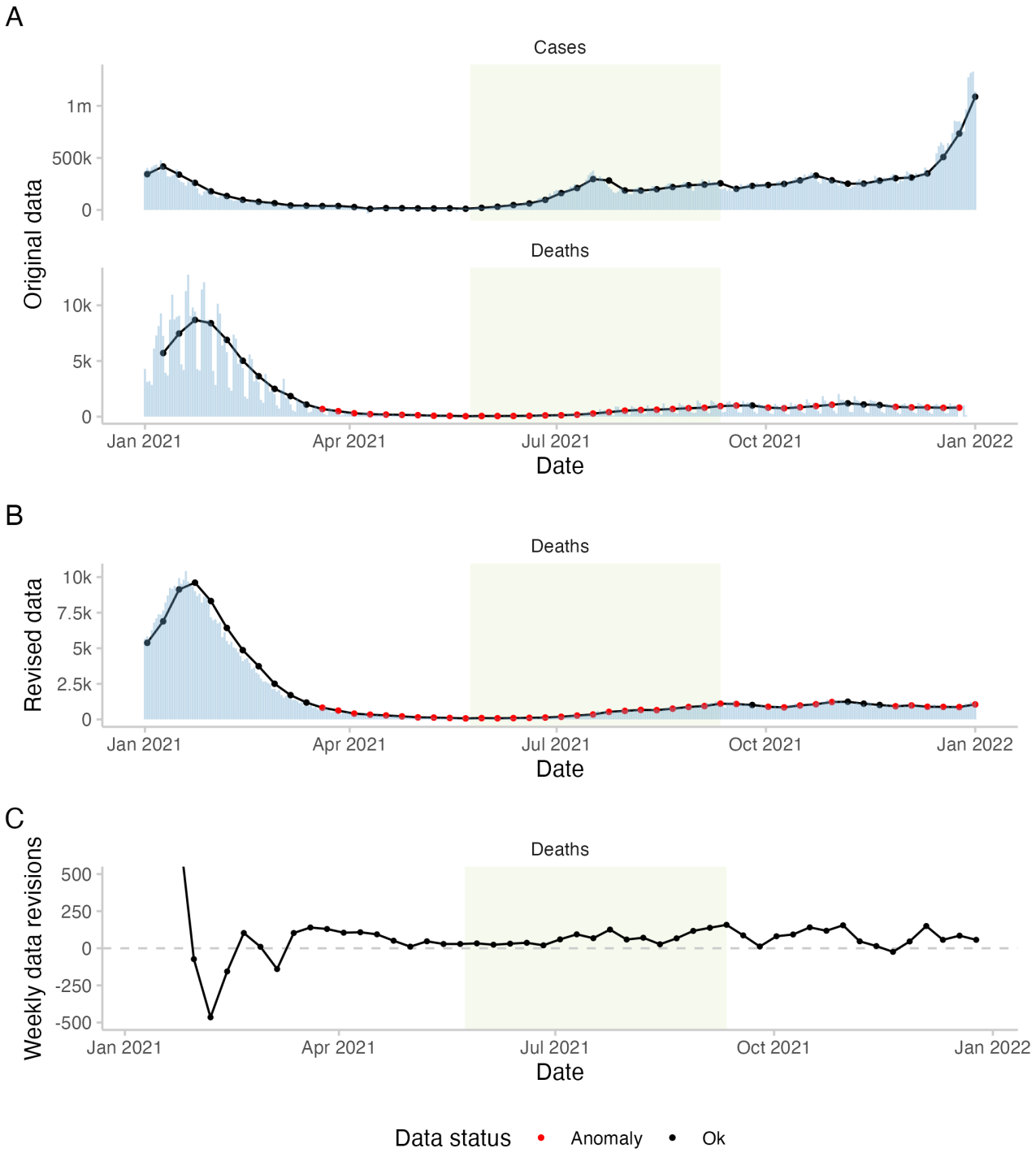

Figure SI.2. **Observed cases and deaths of COVID-19 in the UK. A: Observed daily (bars) and weekly (black lines and points) numbers of cases and deaths as available through the European Forecast Hub when the study concluded in 2021. Daily numbers were multiplied by seven in order to appear on the same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

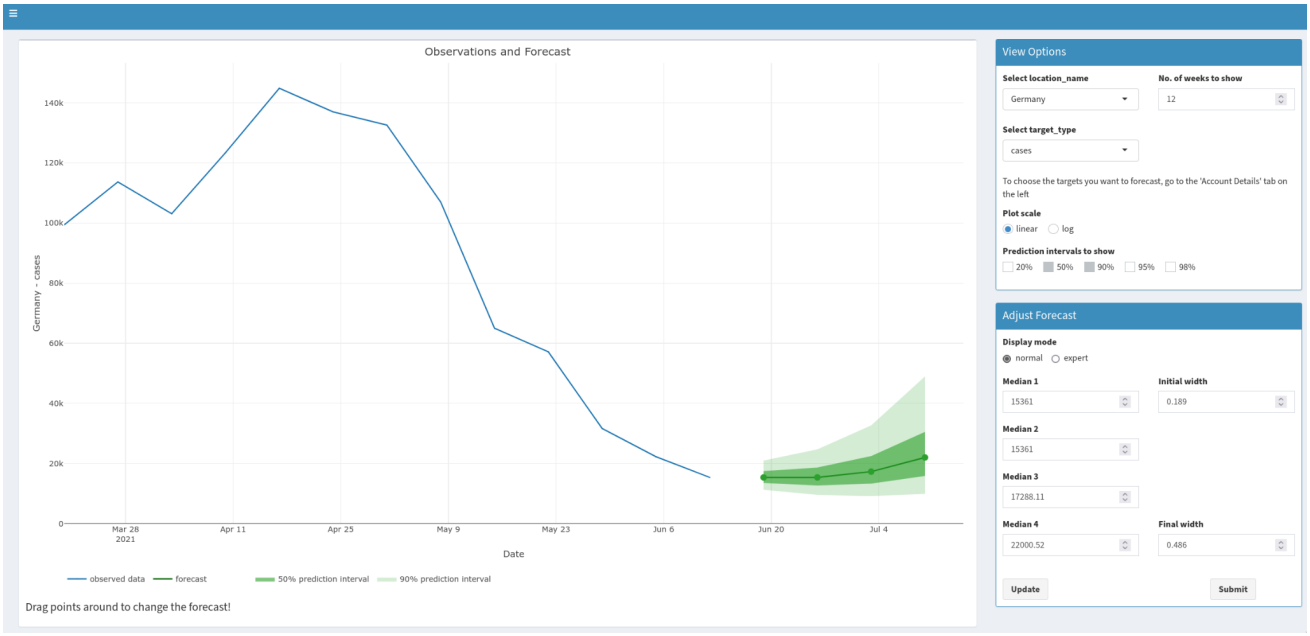

Figure SI.3. Screenshot of the direct forecasting interface.

Figure SI.4. Screenshot of the R_t forecasting interface.

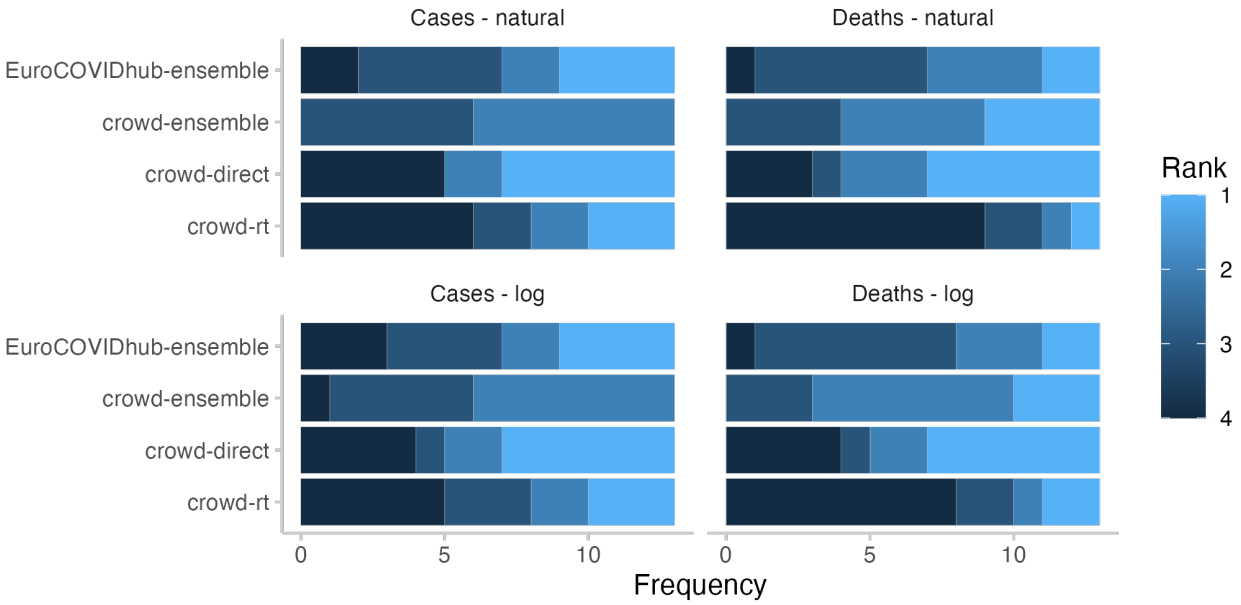

Figure SI.5. Ranks for all forecasting approaches for four week ahead forecasts. Colours indicate how often (out of 13 forecasts) a given approach got 1st, 2nd, 3rd, or 4th rank.

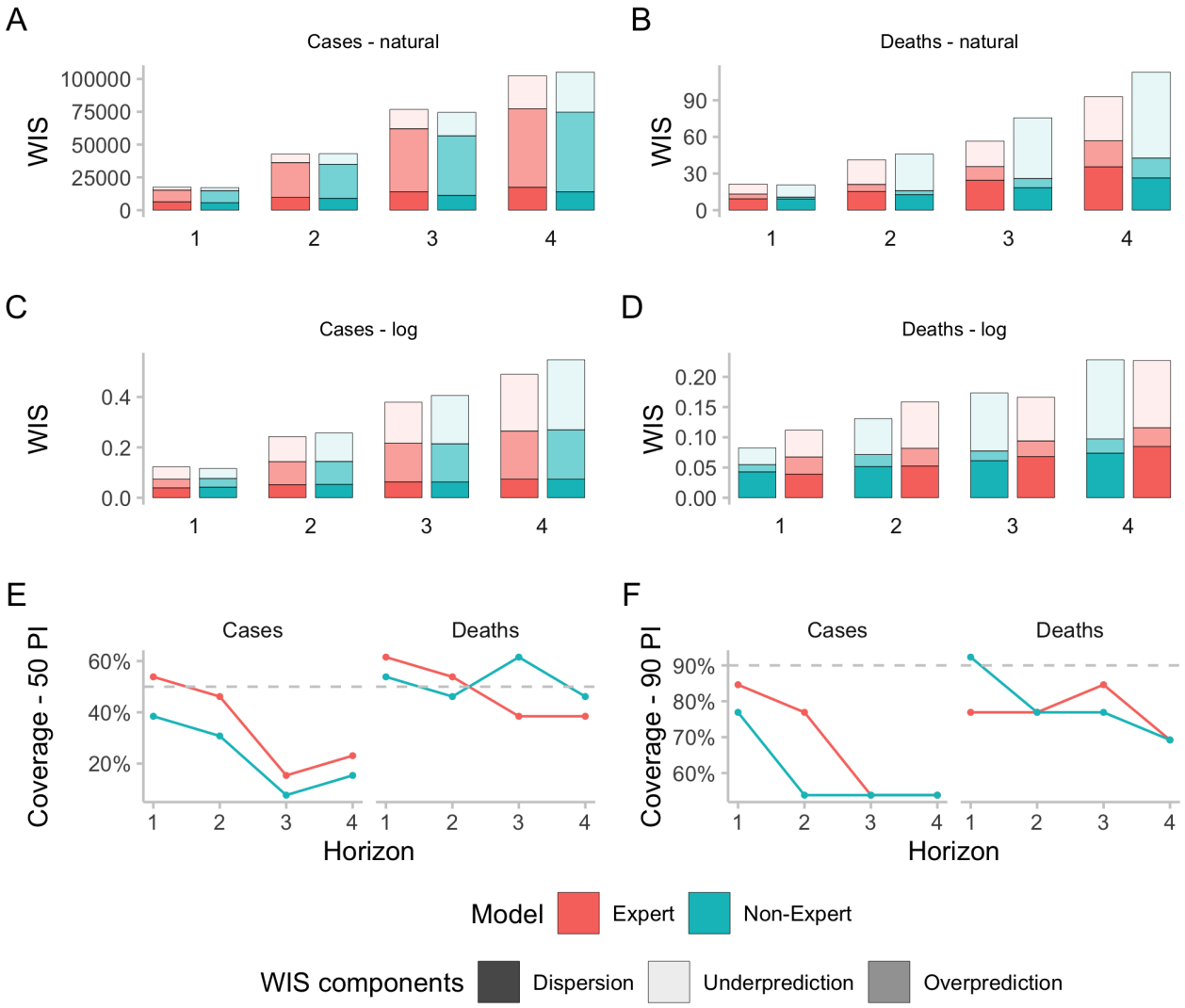

Figure SI.6. **Predictive performance of self-reported "experts" and "non-experts" across forecast horizons. Forecasts from "experts" and "non-experts" were combined to two separate median ensembles, including both direct and R_t forecasts. A-D: WIS stratified by forecast horizon for cases and deaths on the natural and log scale. E, F: Empirical coverage of the 50% and 90% prediction intervals stratified by forecast horizon and target type.**

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

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

Table SI.3. Performance for four-week-ahead forecasts of experts and non-experts. Values have been cut to three significant digits and rounded.

		WIS - natural			WIS - log scale				
Model	Target	abs.	rel.	sd	abs.	rel.	sd	Coverage 50%	Coverage 90%
Expert	Cases	102k	1.04	121k	0.49	1.08	0.4	0.23	0.54
Non-Expert	Cases	105k	1.07	110k	0.55	1.21	0.4	0.15	0.54
crowd-ensemble	Cases	98.4k	1	115k	0.45	1	0.35	0.23	0.62
Expert	Deaths	93	0.95	81.2	0.23	1.17	0.14	0.38	0.69
Non-Expert	Deaths	113	1.15	122	0.23	1.18	0.18	0.46	0.69
crowd-ensemble	Deaths	98.2	1	103	0.19	1	0.13	0.54	0.77

References

[1] Tapiwa Ganyani, Cécile Kremer, Dongxuan Chen, Andrea Torneri, Christel Faes, Jacco Wallinga, and Niel Hens. Estimating the generation interval for coronavirus disease (COVID-19) based on symptom onset data, March 2020. *Euro Surveillance: Bulletin Europeen Sur Les Maladies Transmissibles = European Communicable Disease Bulletin*, 25(17):2000257, April 2020. ISSN 1560-7917. doi: 10.2807/1560-7917.ES.2020.25.17.2000257.

[2] Stephen A. Lauer, Kyra H. Grantz, Qifang Bi, Forrest K. Jones, Qulu Zheng, Hannah R. Meredith, Andrew S. Azman, Nicholas G. Reich, and Justin Lessler. The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application. *Annals of Internal Medicine*, pages M20–0504, March 2020. ISSN 0003-4819. doi: 10.7326/M20-0504.

[3] Bo Xu, Bernardo Gutierrez, Sarah Hill, Samuel Scarpino, Alyssa Loskill, Jessie Wu, Kara Sewalk, Sumiko Mekaru, Alexander Zarebski, Oliver Pybus, David Pigott, and Moritz Kraemer. Epidemiological data from the nCoV-2019 outbreak: Early descriptions from publicly available data. http://virological.org/t/epidemiological-data-from-the-ncov-2019-outbreak-early-descriptions-from-publicly-available-data/337, 2020.

[4] Sam Abbott, Katharine Sherratt, Jonnie Bevan, Hamish Gibbs, Joel Hellewell, James Munday, Patrick Barks, Paul Campbell, Flavio Finger, and Sebastian Funk. Covidregionaldata: Subnational data for the covid-19 outbreak. -, -(-): –, 2020. doi: 10.5281/zenodo.3957539.

[5] Sam Abbott, Joel Hellewell, Katharine Sherratt, Katelyn Gostic, Joe Hickson, Hamada S. Badr, Michael DeWitt, Robin Thompson, EpiForecasts, and Sebastian Funk. *EpiNow2: Estimate Real-Time Case Counts and Time-Varying Epidemiological Parameters*, 2020.

[6] epiforecasts.io/covid. Covid-19: Temporal variation in transmission during the COVID-19 outbreak. https://epiforecasts.io/covid/, 2020.

[7] Katharine Sherratt, Sam Abbott, Sophie R. Meakin, Joel Hellewell, James D. Munday, Nikos Bosse, CMMID Covid-19 working Group, Mark Jit, and Sebastian Funk. Exploring surveillance data biases when estimating the reproduction number: With insights into subpopulation transmission of Covid-19 in England. page 2020.10.18.20214585, March 2021. doi: 10.1101/2020.10.18.20214585.

[8] Stan Development Team. RStan: the R interface to Stan, 2023. URL `https://mc-stan.org/`. R package version 2.21.8.

[9] Gabriel Riutort-Mayol, Paul-Christian Bürkner, Michael R. Andersen, Arno Solin, and Aki Vehtari. Practical Hilbert space approximate Bayesian Gaussian processes for probabilistic programming. *Statistics and Computing*, 33(1): 17, December 2022. ISSN 1573-1375. doi: 10.1007/s11222-022-10167-2.

[10] James A. Scott, Axel Gandy, Swapnil Mishra, Juliette Unwin, Seth Flaxman, and Samir Bhatt. epidemia: Modeling of epidemics using hierarchical bayesian models, 2020. URL `https://imperialcollegelondon.github.io/epidemia/`. R package version 1.0.0.

[11] Samir Bhatt, Neil Ferguson, Seth Flaxman, Axel Gandy, Swapnil Mishra, and James A Scott. Semi-Mechanistic Bayesian modeling of COVID-19 with Renewal Processes. *Journal of the Royal Statistical Society Series A: Statistics in Society*, page qnad030, February 2023. ISSN 0964-1998. doi: 10.1093/jrsssa/qnad030.