Results

Contents

Among our sample of forecasters, the number of forecasts varied over time, as forecasting teams joined or left and contributed to varying combinations of forecast targets. We collated between 7 and 33 models in any one week, forecasting for any combination of 256 possible weekly forecast targets (32 countries, 4 horizons, and 2 target outcomes). Models widely varied in their volume of contributions: on average each model contributed 3869 forecasts, with the median model contributing 764 forecasts.

We categorised 12 models that used human judgement forecasting as qualitative. We further categories 12 models as statistical, 12 as semi-mechanistic, 17 as mechanistic and 3 as agent-based (Supplementary Table). In the volume of forecasts provided, mechanistic, semi-mechanistic, and statistical models each contributed similar numbers of forecasts with approximately one-third each. Qualitative and agent-based models were used much less, with around 1-2% of forecasts represented by them. On average we observed similar performance of the interval score between mechanistic and semi-mechanistic models, performing relatively better than statistical models and worse than the qualitative and agent-based models, although in all these cases with largely overlapping variation in performance. Relative performance among modelling methods also appeared to vary over time (Figure @ref(scores_over_time)). For example, statistical models saw a period of poorer performance over summer 2021, coinciding with the introduction of the Delta variant across Europe.

We considered models forecasting for only one, or multiple countries. We collated 19 single-country models and 28 multi-country models. Single-country models targeted Germany (7 models), Poland (5), Spain (3), Italy (2), Czechia (2) and Slovenia (1). The average multi-country model forecast for a median number of 23 locations. Models classified as targeting multiple countries could vary from week to week in how many locations they forecast. Only 2 models consistently forecast for the same number of locations throughout the entire study period, with 0 of these forecasting for all 32 available locations. On average, multi-country models typically under-performed relative to single-country models to a similar degree over time, although with overlapping range of variation.

Apart from differences between models, we also observed differences based on the epidemiological situation and predictive horizon. Average performance decreased from 1 to 4 weeks of predictive horizon, and was best when the epidemiological situation was stable, with little observed difference between decreasing and increasing scenarios.

We fit a generalised additive mixed model to forecasts' interval scores. The interval score was highly right-skewed with respect to all explanatory variables (see Supplement). We corrected for this by fitting to the log of the interval score. We found no clear evidence that any one type of method structure consistently outperformed others. We also found no evidence for whether the location specificity of the model influenced performance, comparing models forecasting for three or more countries to those targeting only one or two countries.

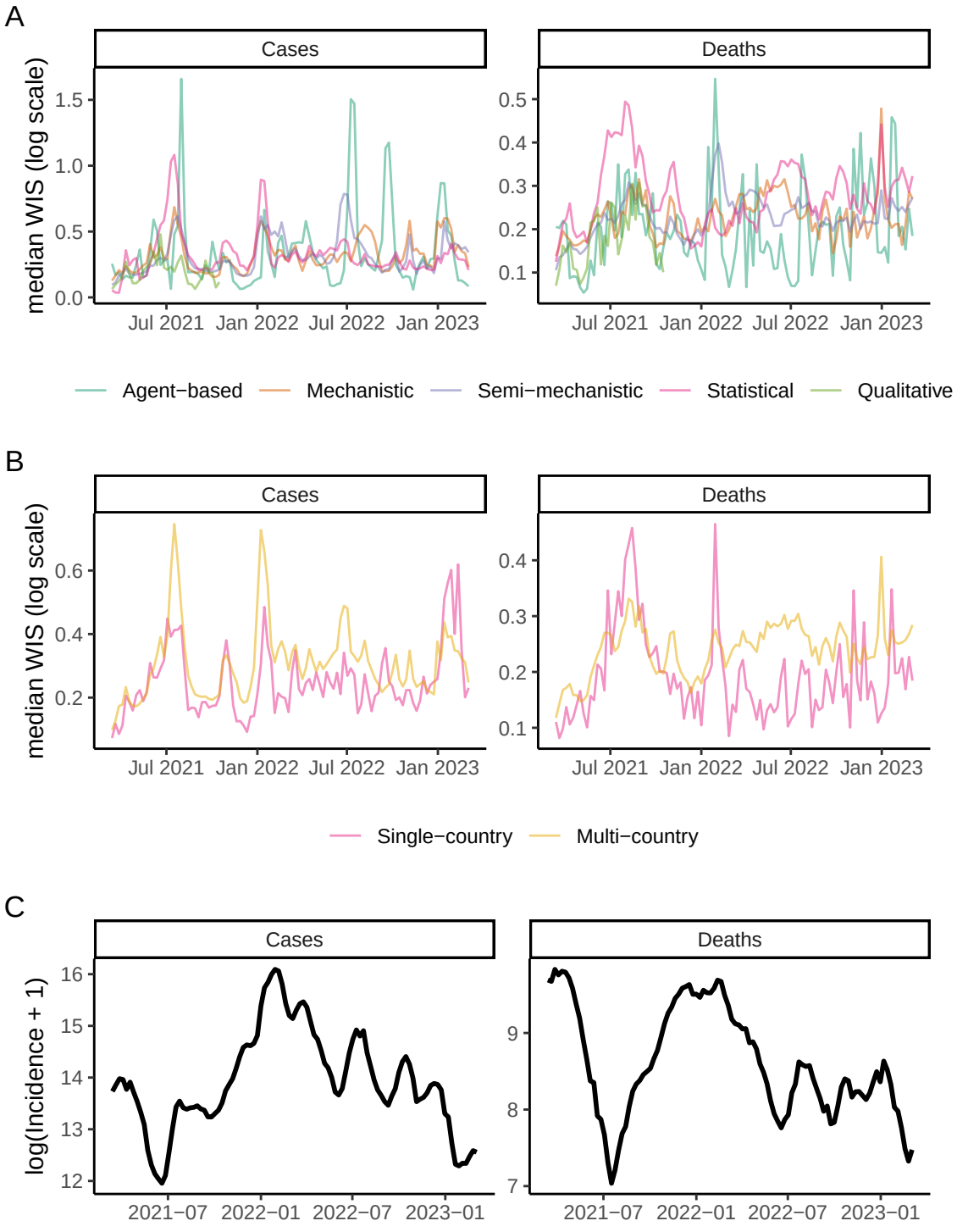

Figure 1: Predictive accuracy of multiple models' forecasts for COVID-19 deaths across Europe. Forecast performance is shown as the median and interquartile range of the weighted interval score, where a lower score indicates better performance. Shown for (A) the Hub ensemble model (the median of all participating forecasts each week); (B) the method used by each model; (C) the number of countries each model targeted (up to 2, or multiple). Forecast performance is summarised across 32 target locations and 1 through 4 week forecast horizons, with varying numbers of forecasters participating over time.

Table 1: Characteristics of forecast performance (interval score) contributed to the European COVID-19 Forecast Hub, March 2021-2023.

	Cases			Deaths		
Variable	Models	Forecasts	Median WIS (IQR)	Models	Forecasts	Median WIS (IQ
Overall	42 (100%)	91966 (100%)	0.28 (0.14-0.63)	38 (100%)	89885 (100%)	0.23 (0.12-0.4
Method						
Agent-based	3 (7.1%)	814 (0.9%)	0.23 (0.11-0.51)	2 (5.3%)	720 (0.8%)	0.17 (0.09-0.3
Mechanistic	16 (38.1%)	27987 (30.4%)	0.28 (0.12-0.64)	16 (42.1%)	33449 (37.2%)	0.22 (0.11-0.4
Semi-mechanistic	9 (21.4%)	28742 (31.3%)	0.27 (0.14-0.62)	10 (26.3%)	28494 (31.7%)	0.22 (0.13-0.3
Statistical	11 (26.2%)	32680 (35.5%)	0.3 (0.15-0.65)	7 (18.4%)	25478 (28.3%)	0.27 (0.14-0.4
Qualitative	3 (7.1%)	1743 (1.9%)	0.18 (0.08-0.42)	3 (7.9%)	1744 (1.9%)	0.15 (0.08-0.3
Number of country targets						
Single-country	19 (45.2%)	3680 (4%)	0.23 (0.1-0.5)	14 (36.8%)	2821 (3.1%)	0.19 (0.09-0.3
Multi-country	23 (54.8%)	88286 (96%)	0.29 (0.14-0.63)	24 (63.2%)	87064 (96.9%)	0.23 (0.13-0.4
Week ahead horizon						
1	42 (100%)	24900 (27.1%)	0.14 (0.07-0.28)	38 (100%)	25417 (28.3%)	0.17 (0.09-0
2	40 (95.2%)	22839 (24.8%)	0.25 (0.13-0.5)	33 (86.8%)	21710 (24.2%)	0.21 (0.12-0.3
3	38 (90.5%)	22247 (24.2%)	0.38 (0.2-0.77)	32 (84.2%)	21498 (23.9%)	0.27 (0.15-0.4
4	37 (88.1%)	21980 (23.9%)	0.52 (0.26-1.05)	31 (81.6%)	21260 (23.7%)	0.34 (0.19-0.6
3-week trend in incidence						
Stable	38 (90.5%)	12682 (13.8%)	0.2 (0.11-0.41)	37 (97.4%)	16793 (18.7%)	0.18 (0.09-0.3
Increasing	41 (97.6%)	36673 (39.9%)	0.31 (0.14-0.74)	37 (97.4%)	32402 (36%)	0.25 (0.13-0.4
Decreasing	42 (100%)	42611 (46.3%)	0.29 (0.15-0.62)	38 (100%)	40690 (45.3%)	0.24 (0.14-0.4

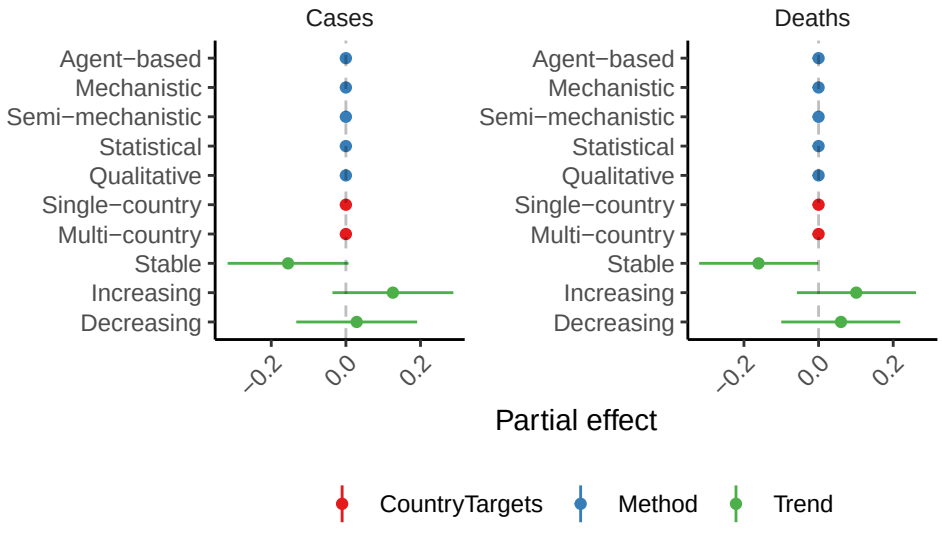

Figure 2: Partial effect size (95% CI) for log-transformed interval score

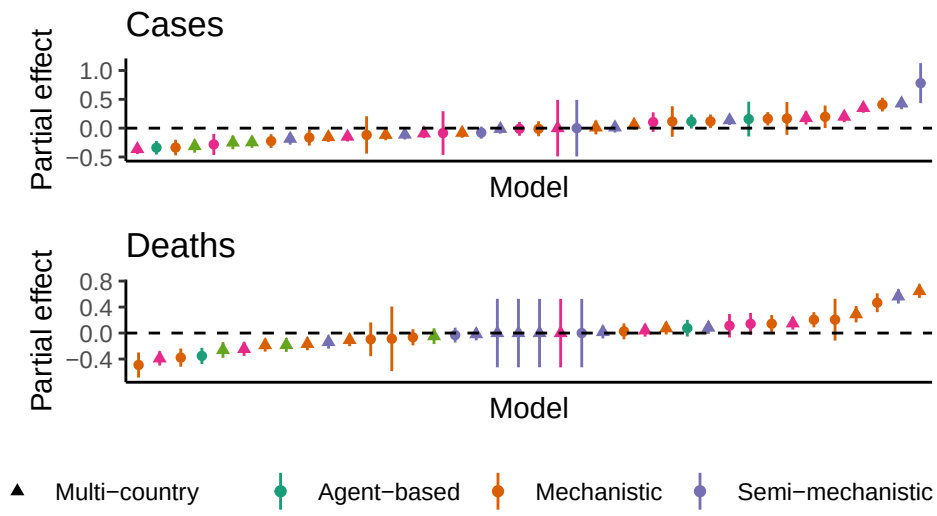

Figure 3: Partial effect size (95% CI) by model