Results

Contents

We evaluated a total 181,851 forecast predictions from 47 forecasting models, contributed by 37 separate modelling teams to the European COVID-19 Forecast Hub (Table 1). 5 teams contributed more than one model. Participating models varied over time as forecasting teams joined or left the Hub and contributed predictions for varying combinations of forecast targets. Between 7 and 33 models contributed in any one week, forecasting for any combination of 256 possible weekly forecast targets (32 countries, 4 horizons, and 2 target outcomes). On average each model contributed 3,869 forecasts, with the median model contributing 764 forecasts.

We categorised 12 models as statistical, 12 as semi-mechanistic, 17 as mechanistic, 3 as agent-based and 3 models that used human judgement forecasting as "other" (Supplementary Table). For 17 (36%) models, investigators disagreed on model classification. The majority of 2/3 was used as the final classification, with additional manual review which in all cases retained the majority decision. In the volume of forecasts provided, mechanistic, semi-mechanistic, and statistical models each contributed similar numbers of forecasts with approximately one-third each. Agent-based and "other" models provided fewer forecasts, representing only 1-2% of forecasts.

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.

We explored the interval score (WIS) as a measure of predictive performance (Figure 1), and characterised its association with model structure and number of countries targeted. We used descriptive statistics and an unadjusted univariate model for each explanatory variable, and then a generalised additive mixed model to give adjusted estimates of the partial effect of each factor while controlling for other sources of variation (Figure 2). The interval score was highly right-skewed with respect to all explanatory variables (see Supplementary Figure 1), which we accounted for by using a log-link.

Descriptively, we noted apparently similar predictive performance between mechanistic, semi-mechanistic, and statistical models. These model structures appeared to perform relatively worse than agent-based and "other" models. For example, in univariate analysis, the partial effect for statistical models forecasting deaths indicates underperformance by 0.165, (95%CI 0.003-0.327) compared to average, while agent-based models performed better than average (-0.24 (-0.43-(-0.05))). However, variation in performance overlapped between all model structures, and we noted relative differences between models may have varied over time (Figure 1). For example, over summer 2021 all model types saw worsening performance coinciding with the introduction of the Delta variant across Europe, but this decline was most marked among statistical models of death outcomes compared to any other model type.

These differences between model structures largely disappeared after adjustment for covariates. We found no clear evidence that any one type of model consistently outperformed others (Figure 2). There was no difference in accuracy between model structures when predicting cases, and we observed only weak differences when predicting deaths. In contrast to unadjusted estimates, we identified that statistical models may have performed slightly better (partial effect -0.03 (-0.16-0.1)), and semi-mechanistic models worse (0.07 (-0.06-0.19)) than the average, although with overlapping uncertainty.

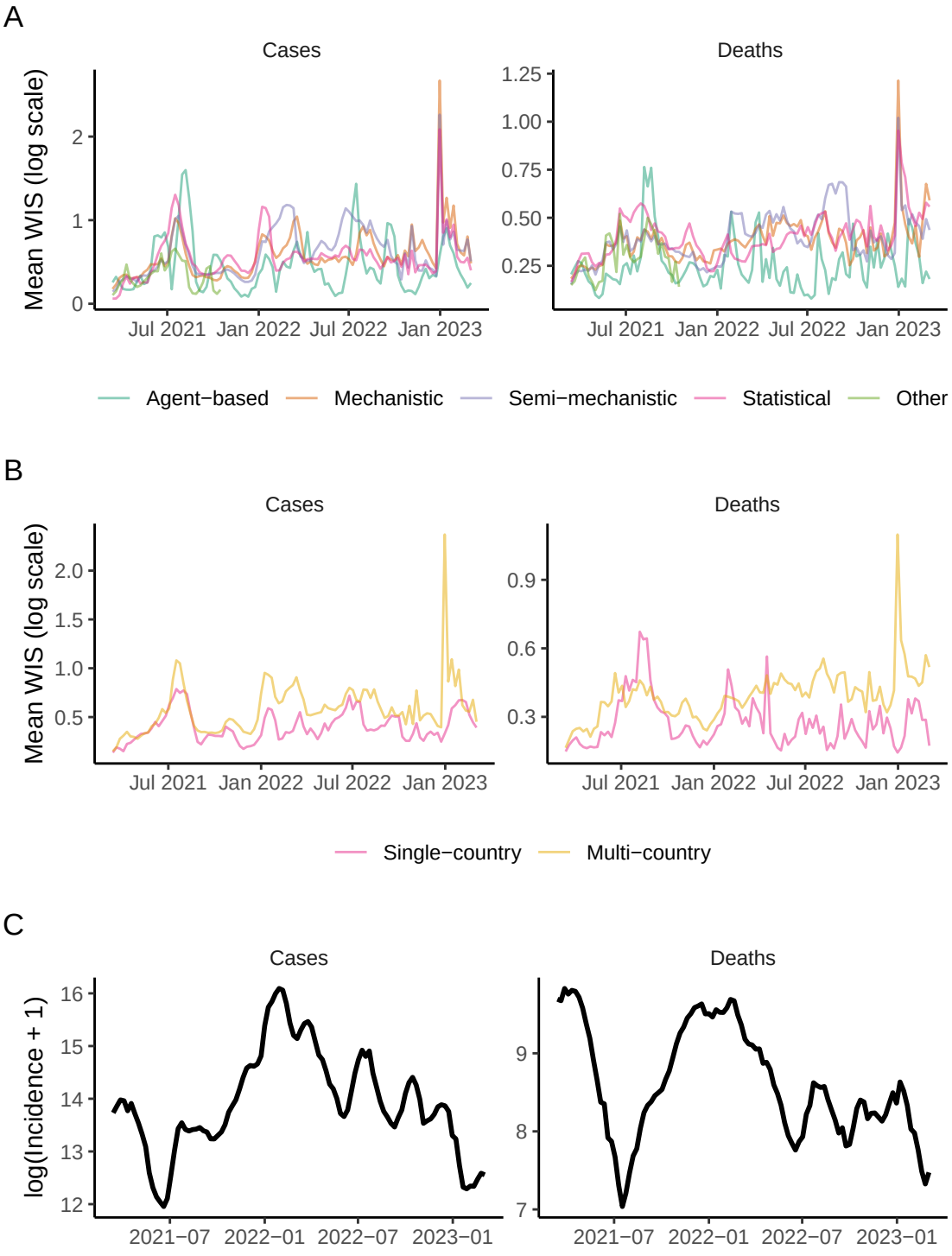

Figure 1: Predictive accuracy of multiple models' forecasts for COVID-19 cases and deaths across 32 European countries over time. Forecast performance is shown as the mean weighted interval score (WIS), where a lower score indicates better performance. Forecast performance is summarised across 32 target locations and 1 through 4 week forecast horizons, with varying numbers of forecasters participating over time. Shown for (A) the method structure used by each model; (B) the number of countries each model targeted (one or multiple); with (C) the total count of observed incidence across all 32 countries, shown on the log scale.

Table 1: Characteristics of forecasts sampled from the European COVID-19 Forecast Hub, March 2021-2023. Forecast performance was measured using the weighted interval score (WIS), with a lower score indicating a more accurate forecast.

	Cases			Deaths		
Variable	Models	Forecasts	Mean WIS (SD)	Models	Forecasts	Mean WIS (SD)
Overall	42 (100%)	91,966 (100%)	0.57 (0.93)	38 (100%)	89,885 (100%)	0.37 (0.51)
Method						
Agent-based	3 (7.1%)	814 (0.9%)	0.44 (0.58)	2 (5.3%)	720 (0.8%)	0.25 (0.25)
Mechanistic	16 (38.1%)	27,987 (30.4%)	0.54 (0.88)	16 (42.1%)	33,449 (37.2%)	0.36 (0.46)
Semi-mechanistic	9 (21.4%)	28,742 (31.3%)	0.59 (1)	10 (26.3%)	28,494 (31.7%)	0.38 (0.6)
Statistical	11 (26.2%)	32,680 (35.5%)	0.58 (0.95)	7 (18.4%)	25,478 (28.3%)	0.39 (0.46)
Other	3 (7.1%)	1,743 (1.9%)	0.33 (0.43)	3 (7.9%)	1,744 (1.9%)	0.29 (0.41)
Number of country targets						
Single-country	19 (45.2%)	3,680 (4%)	0.41 (0.49)	14 (36.8%)	2,821 (3.1%)	0.28 (0.32)
Multi-country	23 (54.8%)	88,286 (96%)	0.57 (0.95)	24 (63.2%)	87,064 (96.9%)	0.38 (0.51)

Considering the number of countries targeted by each model, we descriptively noted that single-country models typically out-performed compared to multi-country models. This relative performance was stable over time, although with overlapping range of variation. Multi-country models appeared to have a more sustained period of poorer performance in forecasting deaths from spring 2022, although we did not observe this difference among case forecasts.

In adjusted estimates, we also saw some indication that models focusing on a single country outperformed those modelling multiple countries (partial effect for single-country models forecasting cases: -0.07 (-0.25-0.11), compared to 0.07 (-0.11-0.25) for multi-country models; and -0.02 (-0.13-0.08) and 0.02 (-0.08-0.13) respectively when forecasting deaths). However, these effects were inconclusive with overlapping uncertainty.

We considered the predictive horizon and epidemiological situation for each forecast as potentially confounding other associations with model performance. In unadjusted estimates, average performance worsened from 1 to 4 weeks of predictive horizon, and was best when the epidemiological situation was stable. Model based analysis supported this observation, albeit with overlapping confidence intervals. This indicated improved predictive performance during stable periods of each country's outbreak curve (cases: -0.24 (-0.57-0.08); deaths: -0.19 (-0.41-0.04)), compared to increasing trends in incidence (cases: -0.07 (-0.39-0.25); deaths: -0.02 (-0.24-0.2)), with worst performance seen when predicting decreasing trends (cases: 0.31 (-0.01-0.64); deaths: 0.2 (-0.02-0.43)).

We identified residual unexplained influences among models' performance. We interpreted the estimated partial random effect for each model as a proxy of its performance whilst correcting for missingness and the common factors considered here. We noted substantial variation beyond the factors we included, indicating that our variable selection was not sufficient for fully explaining performance (Figure 3).

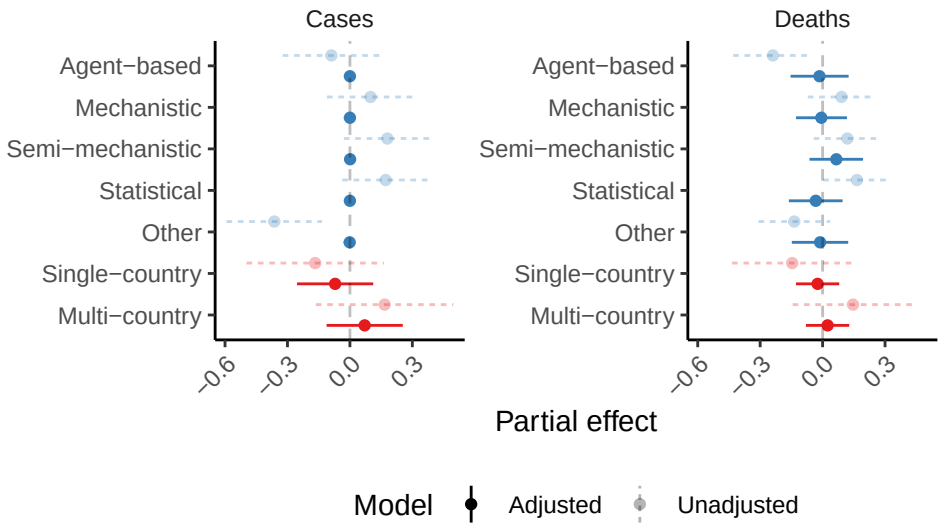

Figure 2: Partial effect (95%CI) on the weighted interval score from model structure and number of countries targeted, before and after adjusting for confounding factors. A lower WIS indicates better forecast performance, meaning effects <0 are relatively better than the group average. Adjusted effects also account for the impact of forecast horizon, epidemic trend, geographic location, and individual model variation. Partial effects and 95% confidence intervals were estimated from fitting a generalised additive mixed model.

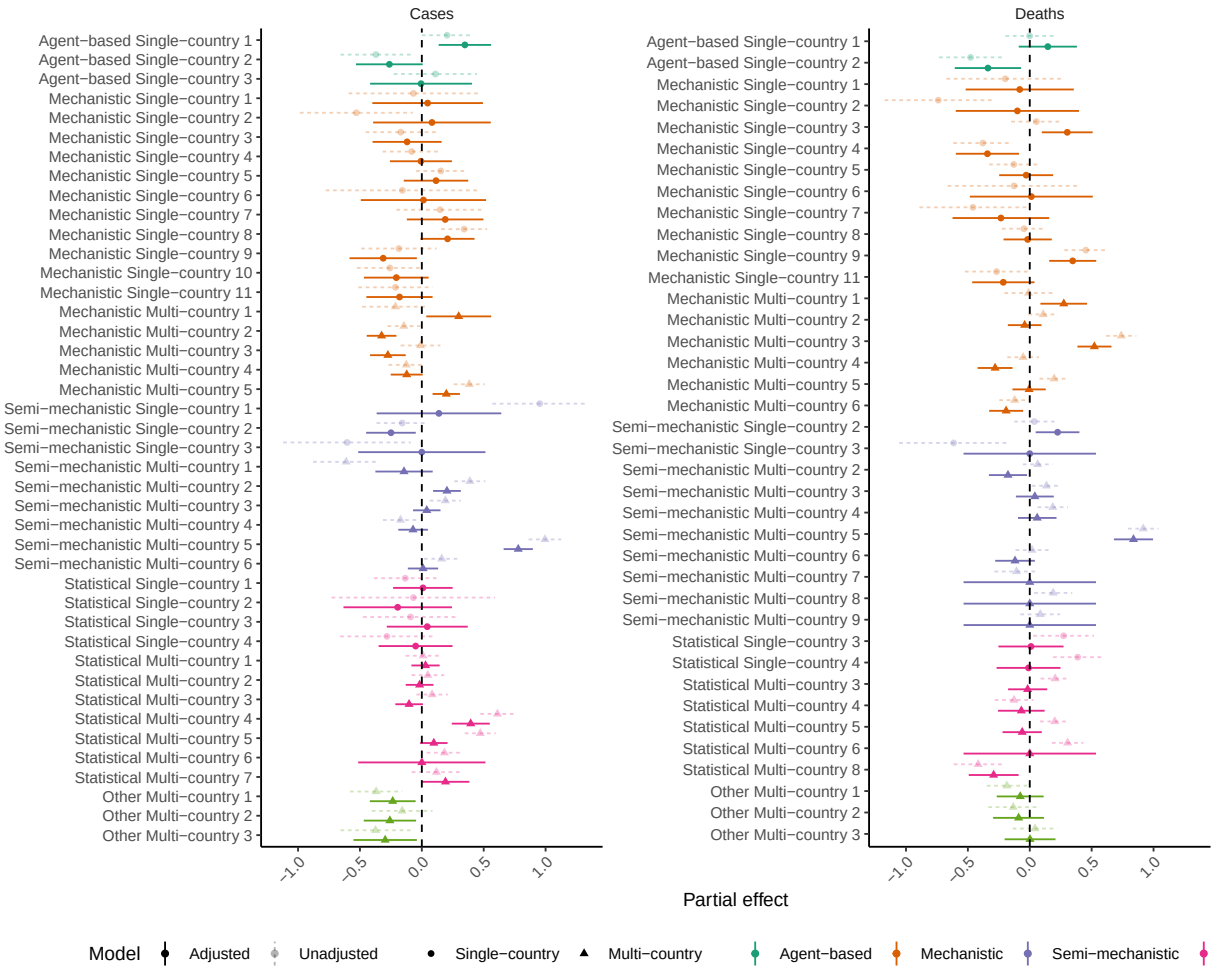

Figure 3: Partial effect size (95% CI) by model. This can be interpreted as adjusted performance after accounting for all other variables in the model, with remaining differences in effects as seen here representing unexplained variation between models beyond these factors.