

Supplement to: Interpreting variation in infectious disease forecast performance with model-based evaluation

Table of contents

Data and participation	1
Source data and code	1
Study participation	2
Covariate derivation	8
Model structure classification	8
Epidemic trend identification	11
Variant phase identification	12
Model specification and fitting	13
Covariate selection	13
Why no separate structure main effect	15
Forecast horizon and individual models	15
Model fitting	15
Estimates	19
Partial effects across covariates	19
Partial effects on the raw scale	21
Model ranking before and after adjustment	24
Individual model effects at one week ahead	26
Sensitivity analyses	26
Data processing	26
Population normalisation	26
Natural scale WIS	27
Model specification	27
Covariate selection	27
Error family and link	28
Structure by outcome under a Gaussian family	28

This supplement supports “Using model-based evaluation to interpret variation in infectious disease forecast performance”. It is organised as the analysis runs: how forecasts and observations were

sourced and who took part; how the covariates were derived; how the model was specified and fitted; the full set of estimates behind the main-text figures; and the sensitivity analyses. Every number and figure here is generated from the same fitted model as the main text, `output/log/results.rds`.

Data and participation

Source data and code

Forecast and observed data were sourced from the European COVID-19 Forecast Hub, available to view at <https://covid19forecasthub.eu/> . All Hub data are now archived at:

- Github: https://github.com/european-modelling-hubs/covid19-forecast-hub-europe__archive
- Zenodo with DOI: <https://doi.org/10.5281/zenodo.13986751>

Data for this work were downloaded on 30th May 2023. These data are available in the Github repository for this paper at: <https://github.com/epiforecasts/eval-by-method/tree/main/data>

The codebase for this paper is publicly available at:

- Github: <https://github.com/epiforecasts/eval-by-method>
- Zenodo with DOI: <https://doi.org/10.5281/zenodo.14903162>

Comments and code contributions are welcome - please use Github [Issues](#).

Please cite code using:

- Katharine Sherratt & Sebastian Funk. (2025). epiforecasts/eval-by-method: Zenodo. <https://doi.org/10.5281/zenodo.14903162>

Study participation

Forecasting teams were recruited to the European Covid-19 Forecast Hub using existing networks and ECDC publicity. Any forecaster was eligible to participate, and there were no selection criteria. All participation was voluntary and unremunerated. Forecasters contributed a standard set of meta-data describing their team and model, and uploaded forecasts weekly. Forecasters were optionally able to express uncertainty by reporting a distribution of up to 23 probabilistic quantiles for each prediction. Forecasts were validated against minimal formatting requirements for quantile intervals and that values were positive integers.

For this study, we collected all forecasts from between 8 March 2021 to 10 March 2023. We excluded forecasts of hospitalisations, which experienced multiple changes in source data during the study period. We excluded forecasts that did not report the full set of 23 quantiles, in order to ensure fair comparison among probabilistic model results. We also excluded the ensemble model created by the Hub team, because it is constructed from the contributed forecasts and so would double-count them. We retained the Hub baseline model: it is an independently specified statistical model, and including it gives a naive reference point against which participant forecasts can be read.

Of 49 models submitting case forecasts during the study period, 43 met these criteria; of 44 submitting death forecasts, 39 did.

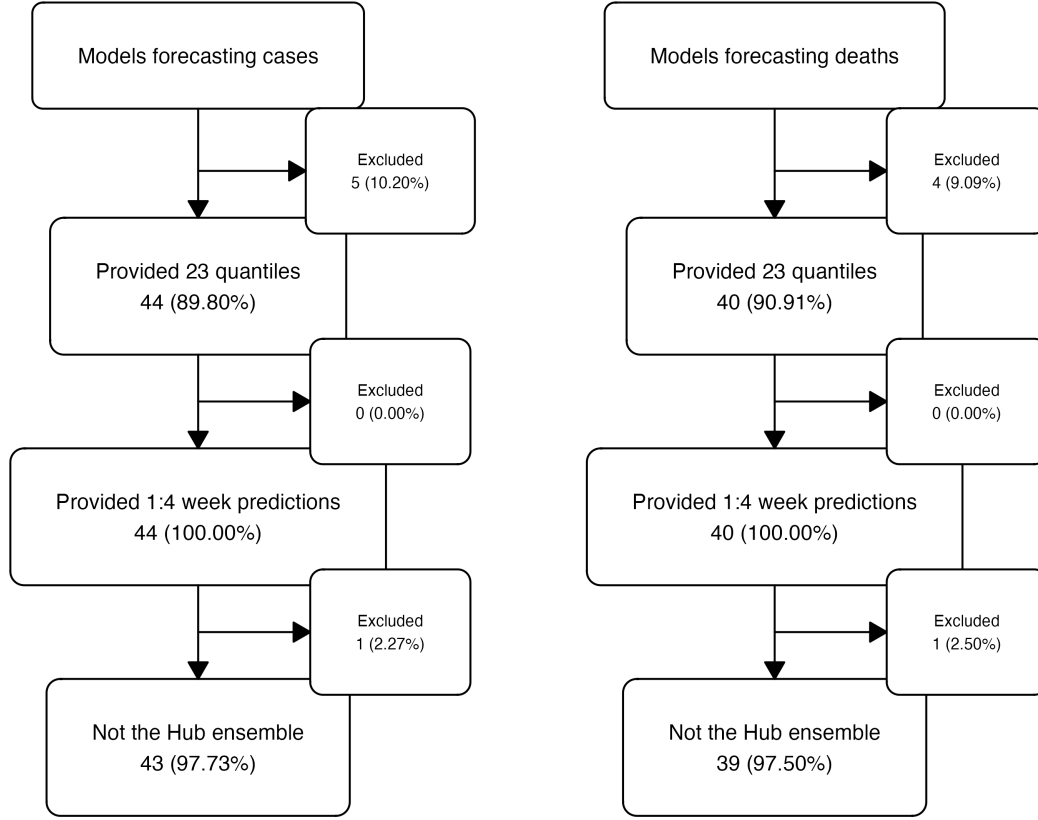


Figure 1: Eligibility criteria for models contributing case (left) and death (right) forecasts to the European COVID-19 Forecast Hub, March 2021 - March 2023

Characteristics of contributing models:

Table 1: Model characteristics contributing to the European COVID-19 Forecast Hub, by method used, number of countries targeted, and number of forecasts contributed.

Model	Method	Country Targets	Case forecasts	Death forecasts
AMM-EpiInvert	Statistical	Multi-country	2,788	
CovidMetrics-epiBATS	Statistical	Single-country	343	
DSMPG-bayes	Semi-mechanistic	Multi-country	760	

Table 1: Model characteristics contributing to the European COVID-19 Forecast Hub, by method used, number of countries targeted, and number of forecasts contributed.

Model	Method	Country Targets	Case forecasts	Death forecasts
EuroCOVIDhub-baseline	Statistical	Multi-country	13,082	13,040
FIAS_FZJ-Epi1Ger	Mechanistic	Single-country	264	264
GoeWroc-BaseBayes	Semi- mechanistic	Single-country	12	
HZI-AgeExtendedSEIR	Mechanistic	Single-country	382	382
ICM-agentModel	Agent-based	Single-country	334	334
IEM_Health-CovidProject	Mechanistic	Multi-country	7,710	7,708
ILM-EKF	Semi- mechanistic	Multi-country	11,998	11,961
ITWW-county_repro	Semi- mechanistic	Multi-country	650	600
Imperial-DeCa	Semi- mechanistic	Multi-country		571
Imperial-RtI0	Semi- mechanistic	Multi-country		571
Imperial-sbcp	Semi- mechanistic	Multi-country		571
JBUD-HMXK	Mechanistic	Multi-country	1,324	1,324
KITmetricslab- bivar_branching	Statistical	Single-country	8	
Karlen-pypm	Mechanistic	Multi-country	3,208	3,186
LANL-GrowthRate	Semi- mechanistic	Multi-country	3,692	3,696
LeipzigIMISE-SECIR	Mechanistic	Single-country	16	16
MIMUW-StochSEIR	Mechanistic	Single-country	76	76
MIT_CovidAnalytics- DELPHI	Mechanistic	Multi-country	348	500
MOCOS-agent1	Agent-based	Single-country	386	386
MUNI-ARIMA	Statistical	Multi-country	10,979	11,314
MUNI-LaggedRegARIMA	Statistical	Multi-country		736
MUNI-VAR	Statistical	Multi-country	976	976
MUNI_DMS-SEIAR	Mechanistic	Single-country	224	200
PL_GRedlarski- DistrictsSum	Mechanistic	Single-country	378	
RobertWalraven-ESG	Statistical	Multi-country	9,190	10,465
SDSC_ISG-TrendModel	Statistical	Multi-country	1,756	1,744
UB-BSLCoV	Statistical	Single-country	96	96
UC3M-EpiGraph	Agent-based	Single-country	94	
ULZF-SEIRC19SI	Mechanistic	Single-country	249	249
UMass-MechBayes	Mechanistic	Multi-country		5,948
UMass-SemiMech	Semi- mechanistic	Multi-country	1,888	1,904

Table 1: Model characteristics contributing to the European COVID-19 Forecast Hub, by method used, number of countries targeted, and number of forecasts contributed.

Model	Method	Country Targets	Case forecasts	Death forecasts
UNED-PreCoV2	Statistical	Single-country	147	147
UNIPV-BayesINGARCHX	Statistical	Multi-country	426	
USC-SIkJalpha	Mechanistic	Multi-country	12,900	12,688
UpgUmibUsi-MultiBayes	Semi- mechanistic	Single-country	99	99
bisop-seirfilter	Mechanistic	Single-country	32	32
bisop-seirfilterlite	Mechanistic	Multi-country	336	336
epiMOX-SUIHTER	Mechanistic	Single-country	134	134
epiforecasts-EpiExpert	Judgement	Multi-country	945	948
epiforecasts-EpiExpert_Rt	Judgement	Multi-country	404	404
epiforecasts-EpiExpert_direct	Judgement	Multi-country	394	392
epiforecasts-EpiNow2	Semi- mechanistic	Multi-country	8,843	7,721
epiforecasts-weeklygrowth	Statistical	Multi-country	5,971	
itwm-dSEIR	Mechanistic	Single-country	406	406
prolix-euclidean	Semi- mechanistic	Multi-country	800	800

Number of models participating in forecasting 1-week ahead case incidence for each country over the study period.

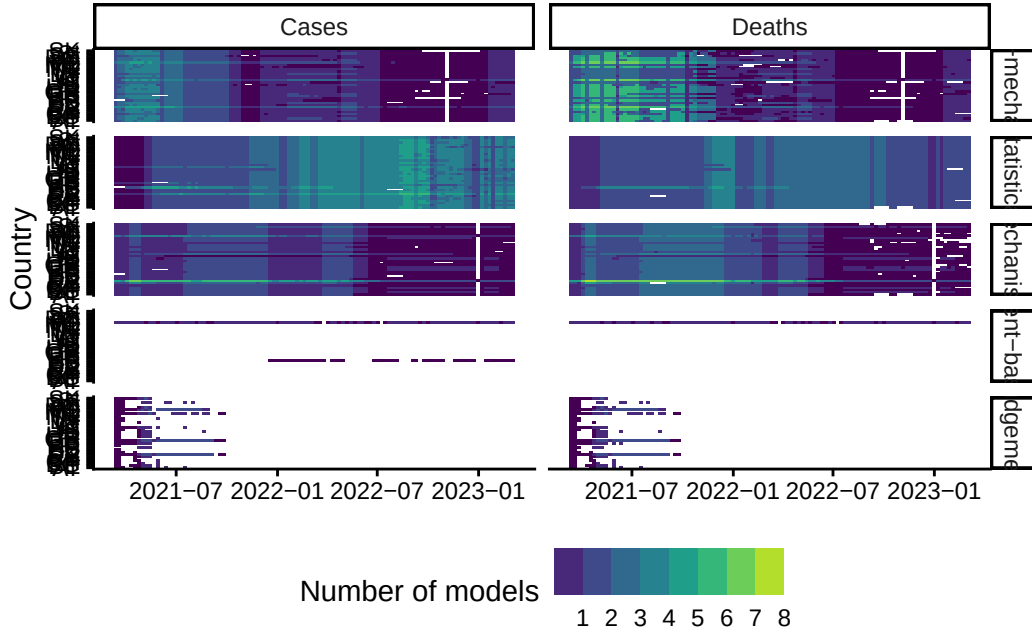


Figure 2: Participation varied by country and over time, and by model structure. Number of models forecasting one week ahead for each country and week, by model structure and epidemiological outcome.

We explored how models selected geographic targets over time. Forecasters both added and removed targets among the set for which they forecast each week. This figure shows number of targets submitted by each model at the 1-week ahead horizon. We noted the same variation for 2-4 week forecasts.

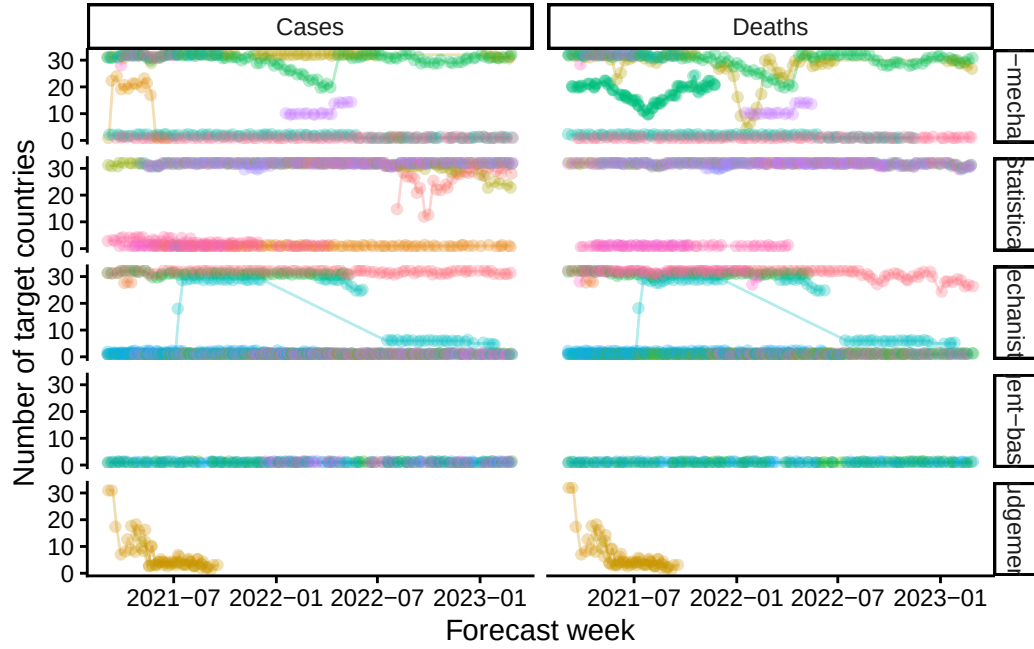


Figure 3: The number of countries targeted by each model forecasting for 1-week ahead incidence of Covid-19 cases or deaths, shown by classification of model structure. Each forecaster selected any number of targets from a set of 32 countries.

The main text shows this same relationship as aggregated percentile bands. At the level of individual forecasts, the same data show where forecasts actually concentrate and where scores cluster or trail off.

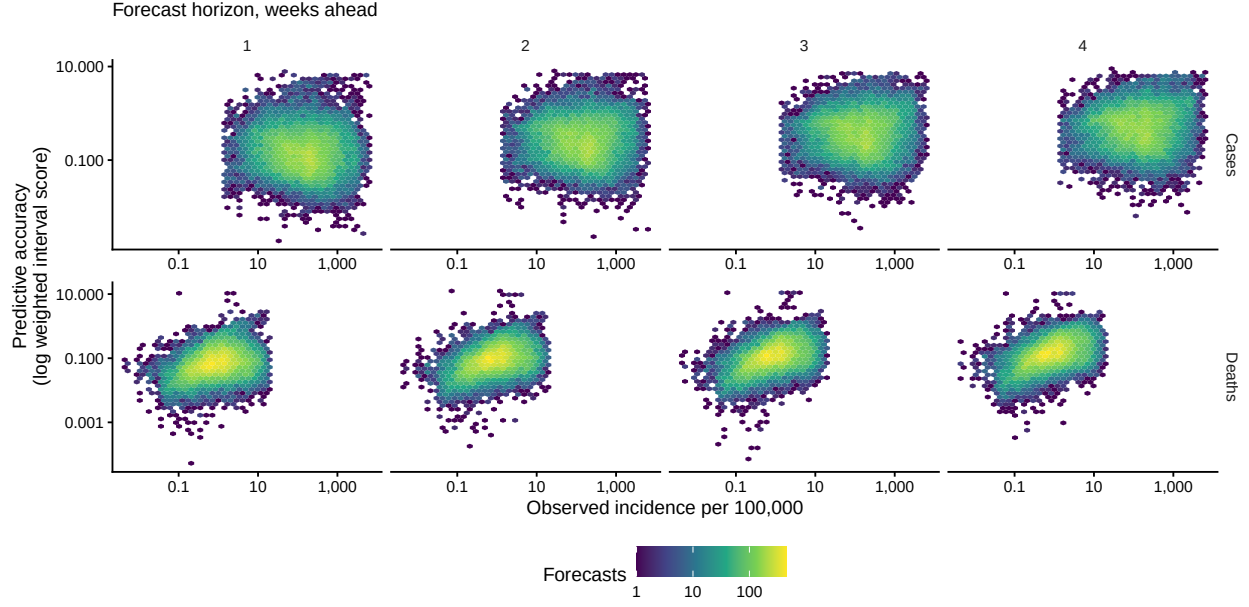


Figure 4: Density of individual forecasts by observed incidence per 100,000 population and weighted interval score (WIS, log scale), for COVID-19 cases and deaths across 32 European countries, faceted by forecast horizon (columns, weeks ahead). Hexagonal bins are coloured by the number of forecasts they contain (log scale). This is the forecast-level detail behind the main text’s percentile bands of the same relationship.

Covariate derivation

Model structure classification

Human raters classified models according to methodological structures.

Table 2: Classification of models by number of raters in total and agreement on model structure.

Model	Final classification	Agreement	Total raters	Semi-mechanistic	Mechanistic	Agent-based	Statistical	Judgement	Other	Machine Learning
AMM-EpiInvert	Statistical	FALSE	3	0	0	0	2	0	1	0
DSMPG-bayes	Semi-mechanistic	FALSE	3	2	0	0	1	0	0	0
ILM-EKF	Semi-mechanistic	FALSE	3	2	0	0	1	0	0	0
ITWW-county_repro	Semi-mechanistic	FALSE	3	2	0	0	1	0	0	0
Imperial-DeCa	Semi-mechanistic	FALSE	3	2	0	0	1	0	0	0

Table 2: Classification of models by number of raters in total and agreement on model structure.

Model	Final classifica- tion	Agreement	Total raters	Semi- mechanistic	Mechanistic	Agent- based	Statistical	Judgment	Other	Machine Learn- ing
Imperial-RtI0	Semi- mechanistic	FALSE	3	2	0	0	1	0	0	0
Imperial- sbkp	Semi- mechanistic	FALSE	3	2	0	0	1	0	0	0
KITmetricslab- bivar_branching	Statistical	FALSE	3	1	0	0	2	0	0	0
Karlen-pypm	Mechanistic	FALSE	3	0	2	0	1	0	0	0
LANL- GrowthRate	Semi- mechanistic	FALSE	3	2	0	0	1	0	0	0
SDSC_ISG- TrendModel	Statistical	FALSE	3	0	0	0	2	0	1	0
UMass- SemiMech	Semi- mechanistic	FALSE	4	3	1	0	0	0	0	0
USC- SlkJalpha	Mechanistic	FALSE	4	1	3	0	0	0	0	0
UpgUmibUsi- MultiBayes	Semi- mechanistic	FALSE	3	2	0	0	1	0	0	0
bisop- seirfilter	Mechanistic	FALSE	4	1	3	0	0	0	0	0
bisop- seirfilterlite	Mechanistic	FALSE	4	1	3	0	0	0	0	0
prolix- euclidean	Semi- mechanistic	FALSE	4	3	0	0	0	0	1	0
CovidMetrics- epiBATS	Statistical	TRUE	3	0	0	0	3	0	0	0
EuroCOVIDhub baseline	Statistical	TRUE	4	0	0	0	4	0	0	0
FIAS_FZJ- Epi1Ger	Mechanistic	TRUE	3	0	3	0	0	0	0	0
GoeWroc- BaseBayes	Semi- mechanistic	TRUE	3	3	0	0	0	0	0	0
HZI- AgeExtendedSEIR	Mechanistic	TRUE	3	0	3	0	0	0	0	0
ICM- agentModel	Agent- based	TRUE	3	0	0	3	0	0	0	0
IEM_Health- CovidProject	Mechanistic	TRUE	4	0	4	0	0	0	0	0
JBUD- HMXK	Mechanistic	TRUE	3	0	3	0	0	0	0	0
LeipzigIMISE- SECIR	Mechanistic	TRUE	3	0	3	0	0	0	0	0
MIMUW- StochSEIR	Mechanistic	TRUE	3	0	3	0	0	0	0	0

Table 2: Classification of models by number of raters in total and agreement on model structure.

Model	Final classifica- tion	Agreement	Total raters	Semi- mechanistic	Mechanistic	Agent- based	Statistical	Judgement	Other	Machine Learn- ing
MIT_CovidAnalytics	Mechanistic	TRUE	3	0	3	0	0	0	0	0
DELPHI										
MOCOS-agent1	Agent- based	TRUE	3	0	0	3	0	0	0	0
MUNI-ARIMA	Statistical	TRUE	3	0	0	0	3	0	0	0
MUNI-LaggedRegARIMA	Statistical	TRUE	3	0	0	0	3	0	0	0
MUNI-VAR	Statistical	TRUE	3	0	0	0	3	0	0	0
MUNI_DMS-SEIAR	Mechanistic	TRUE	3	0	3	0	0	0	0	0
PL_GRedlarski-DistrictsSum	Mechanistic	TRUE	3	0	3	0	0	0	0	0
RobertWalravenESG	Statistical	TRUE	3	0	0	0	3	0	0	0
UB-BSLCoV	Statistical	TRUE	3	0	0	0	3	0	0	0
UC3M-EpiGraph	Agent- based	TRUE	3	0	0	3	0	0	0	0
ULZF-SEIRC19SI	Mechanistic	TRUE	3	0	3	0	0	0	0	0
UMass-MechBayes	Mechanistic	TRUE	3	0	3	0	0	0	0	0
UNED-PreCoV2	Statistical	TRUE	3	0	0	0	3	0	0	0
UNIPV-BayesINGARCHX	Statistical	TRUE	3	0	0	0	3	0	0	0
epiMOX-SUIHTER	Mechanistic	TRUE	3	0	3	0	0	0	0	0
epiforecasts-EpiExpert	Judgement	TRUE	3	0	0	0	0	3	0	0
epiforecasts-EpiExpert_Rt	Judgement	TRUE	3	0	0	0	0	3	0	0
epiforecasts-EpiExpert_direct	Judgement	TRUE	3	0	0	0	0	3	0	0
epiforecasts-EpiNow2	Semi- mechanistic	TRUE	3	3	0	0	0	0	0	0
epiforecasts-weeklygrowth	Statistical	TRUE	2	0	0	0	2	0	0	0
itwm-dSEIR	Mechanistic	TRUE	3	0	3	0	0	0	0	0

Epidemic trend identification

We categorised each week as “Stable”, “Decreasing”, or “Increasing”, based on the difference over a three-week moving average of incidence (with a change of $\pm 5\%$ as “Stable”).

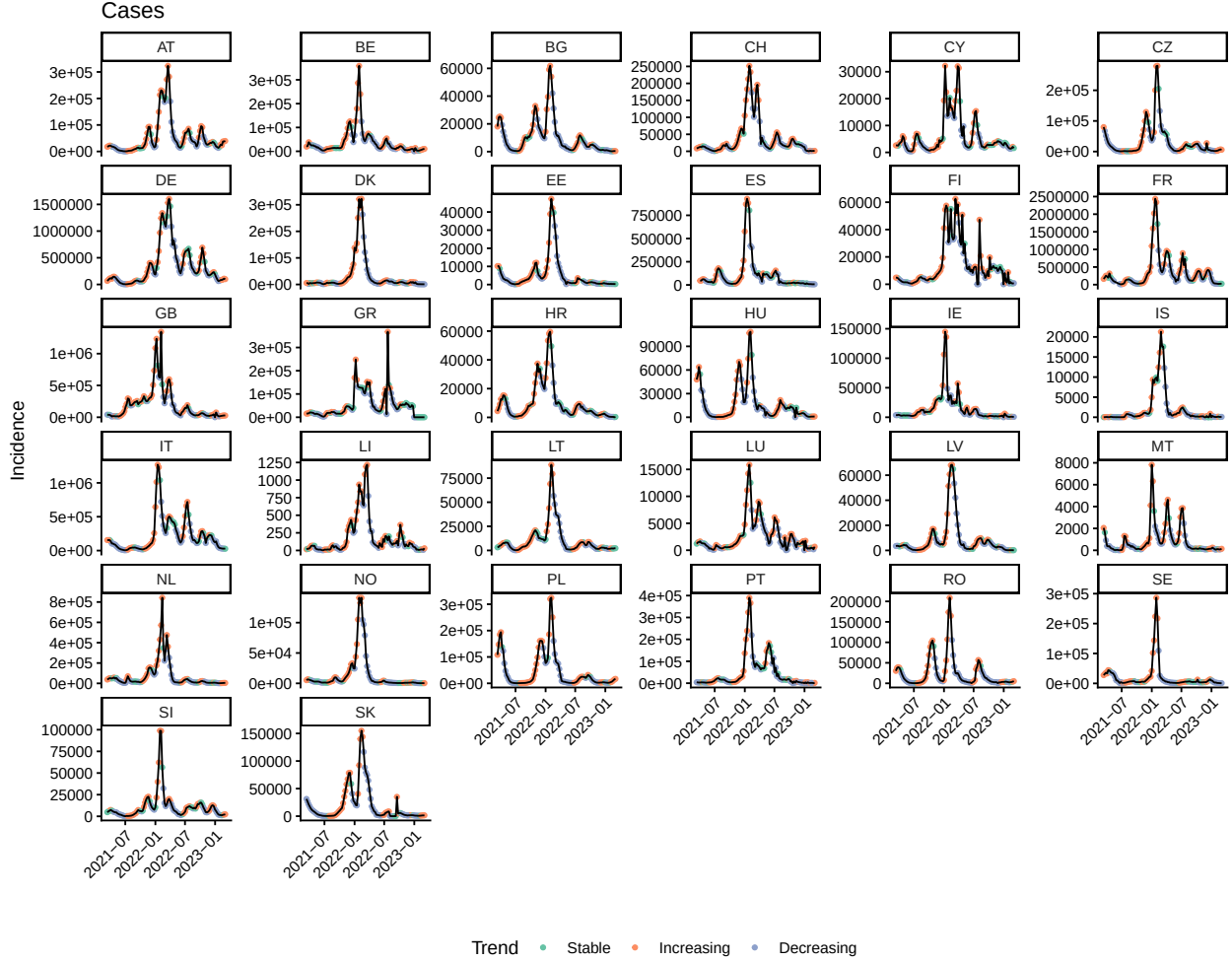


Figure 5: Weekly incidence of reported cases in each country, coloured by the epidemic trend assigned to that week (“Stable”, “Increasing” or “Decreasing”), from the change in a three-week moving average.

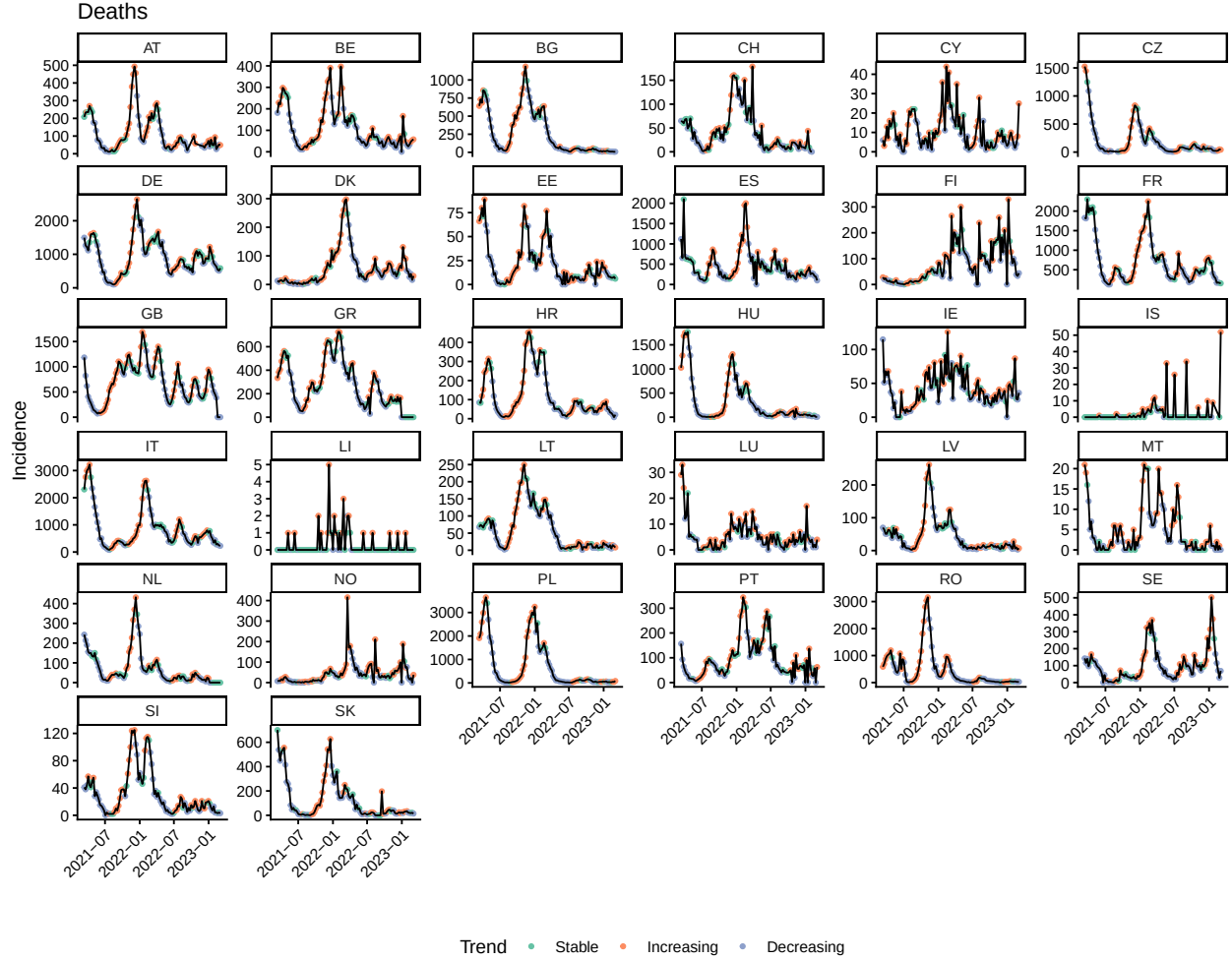


Figure 6: Weekly incidence of reported deaths in each country, coloured by the epidemic trend assigned to that week, derived as for cases.

Variant phase identification

Genomic surveillance data were obtained from three sources: ECDC (covering 30 European countries), UKHSA (Great Britain), and the Swiss Federal Office of Public Health (Switzerland). Variant lineages were mapped to six named phases in expected chronological order: Alpha, Delta, Omicron-BA.1, Omicron-BA.2, Omicron-BA.4/5, and Omicron-BQ/XBB. For each country, we identified the first week in which each named variant exceeded 50% of sequenced samples. We enforced chronological ordering by removing any out-of-sequence phases, then expanded phase assignments to all weeks by filling forward and backward from observed transition dates. This per-location approach accounts for the fact that variant dominance dates differed substantially across European countries. Where genomic surveillance data were too sparse to identify a transition (Hungary), we supplemented with epidemiological reports to set the Alpha-to-Delta transition date.

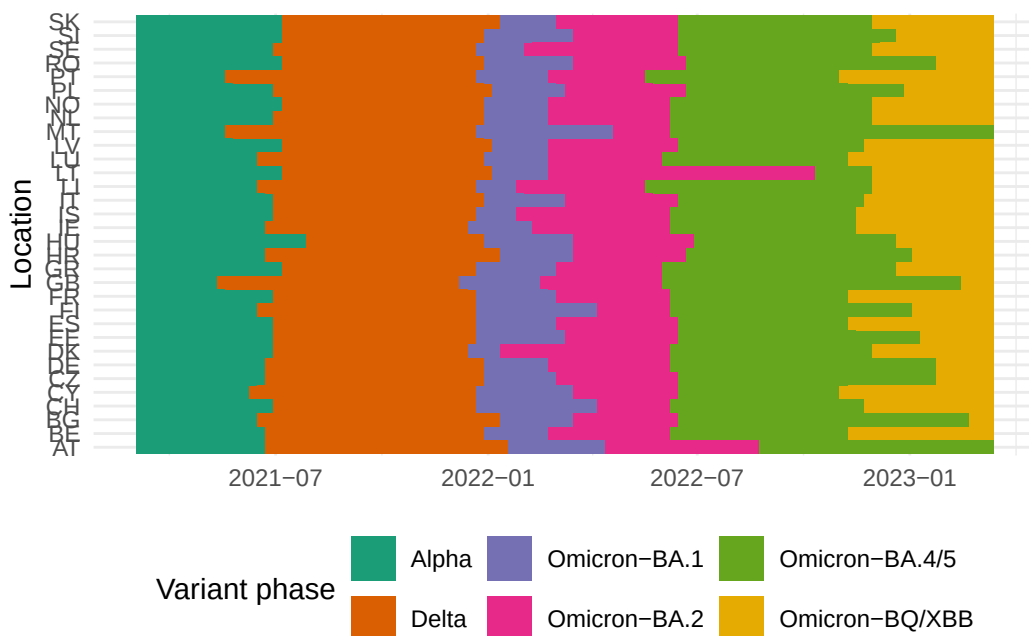


Figure 7: Variant phases identified by dominant variant in each location and week

Model specification and fitting

We developed this work in two major versions. The first aimed to evaluate both model structure and geographic target specificity (one or multiple countries) and their effect on model performance. The second substantially revised the aim and scope:

- Limited aim to developing the use of a modelling approach to evaluate forecast performance
- Specified a single response (model structure) and treating other covariates as confounding factors
- Improved range of confounding factors (epidemiological outcomes covering both cases and deaths; adding a covariate for dominant variant phase)
- More extensive model checking and diagnostics to address highly skewed residuals

See the [sensitivity analyses](#) below for alternative model specifications.

Covariate selection

We aimed to assess the direct relationship between Method (i.e. model structure), and LWIS. We show the assumptions underlying our strategy in a causal diagram. We emphasise that model specification is extremely flexible, and we approach this work is exploratory rather than fully inferential.

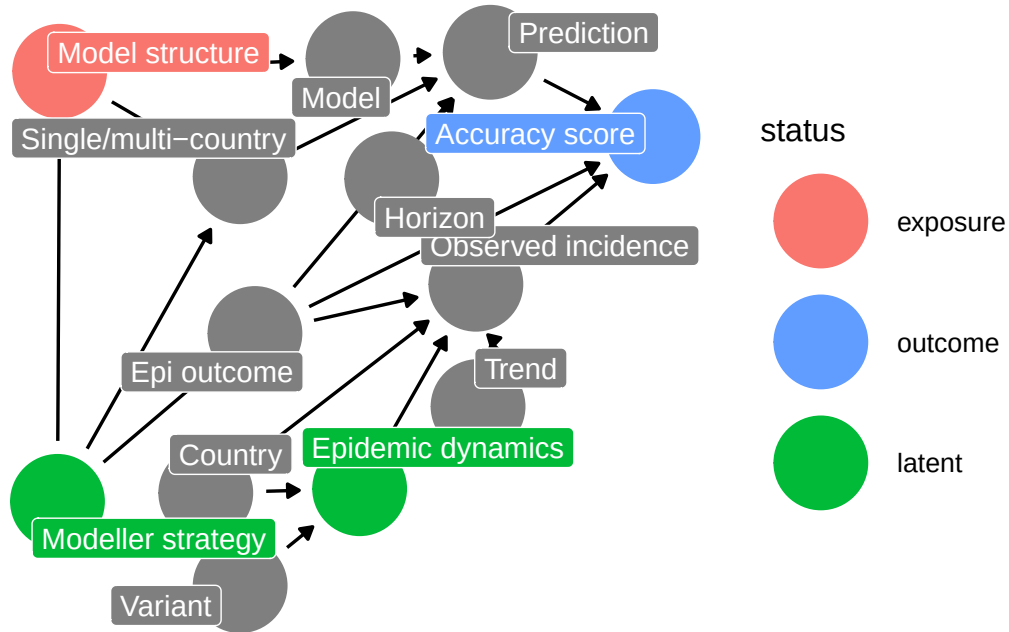


Figure 8: Directed acyclic graph of assumed causal relationships among forecast performance (wis), model structure (Method), and covariates.

The epidemiological outcome enters the diagram as a confounder rather than simply a covariate. Forecasters chose which outcomes to submit predictions for, and that choice is associated with model structure, opening a backdoor path from structure to the score through modeller strategy. The outcome also changes the scale and reporting characteristics of the series being scored.

The joint model therefore adjusts for:

- Forecast target difficulty (which block backdoor paths from epidemic dynamics into the score), as:
 - EpiTarget, Trend, Horizon, Incidence, Location, VariantPhase
- Forecaster strategy, as:
 - CountryTargets, Model

Querying the diagram for a minimal sufficient adjustment set returns this exact set of covariates, for the direct effect of model structure on the score. It returns no valid adjustment set for the total effect: modeller strategy is unobserved and confounds model structure with the choice of geographic scope, so no set of measured covariates blocks that path. This is why we report a partial, direct association throughout rather than a total effect, and why residual confounding by unmeasured forecaster characteristics remains a limitation.

Adjustment gives the partial association between model structure and LWIS with these covariates held fixed at their observed values.

The joint model also fits an interaction between model structure and the epidemiological outcome, allowing a structure to predict one outcome relatively better than the other. This is effect modification rather than confounding. A causal diagram encodes conditional independence and not effect

modification, so the interaction does not appear in the diagram and does not change the adjustment set.

Why no separate structure main effect

The crossed term `s(Method, Epi_target, bs = "re")` is an unconstrained zero-mean effect over all structure-by-outcome cells, so its average across outcomes is what a structure main effect would represent. With both terms in the model and both penalised, the division between them follows the relative variance estimates rather than the data: fitted together, `mgcv` gave the main effect 0.001 effective degrees of freedom against 4.9 for the crossed term, and dropping it left AIC, deviance explained and residuals unchanged. We therefore fit the crossed term alone and recover the pooled per-structure effect as a contrast averaging a structure's two cells.

The same aliasing applies to the epidemiological outcome, whose main effect is the average of the crossed effects within an outcome. Here only one of the two terms is penalised, so the unpenalised fixed effect takes the component common to all structures and the crossed term keeps only departures from it. In the fitted model the crossed effects average to zero within each outcome to numerical precision, so the whole deaths-versus-cases difference sits in the fixed coefficient. This is emergent rather than imposed: the smooth retains all ten coefficients, so no centring constraint was applied. We did not treat the outcome as a random effect, which with two levels gives no basis for estimating a variance and would shrink a large, well-identified contrast toward zero.

Forecast horizon and individual models

Performance is known to worsen with horizon, so the model includes one horizon curve shared by all models, `s(Horizon, k = 4)`, and a factor smooth giving each model its own deviation from it, `s(Horizon, Model, k = 4, bs = "fs")`. The factor smooth penalises each model's whole deviation, including its level and slope, towards zero. It therefore also carries the individual model effect, and there is no separate `s(Model, bs = "re")`: with both terms, a model's level would be split between them by the relative variance estimates. We report a model's effect as its deviation averaged over the horizons it forecast, so no model is judged on horizons it did not forecast. The same deviation at one week ahead, which every model forecast, is given under [Estimates](#).

Alternative specifications are compared in `R/sensitivity/check-horizon-spec.qmd`.

Model fitting

We noted our outcome of LWIS had a strongly skewed PDF.

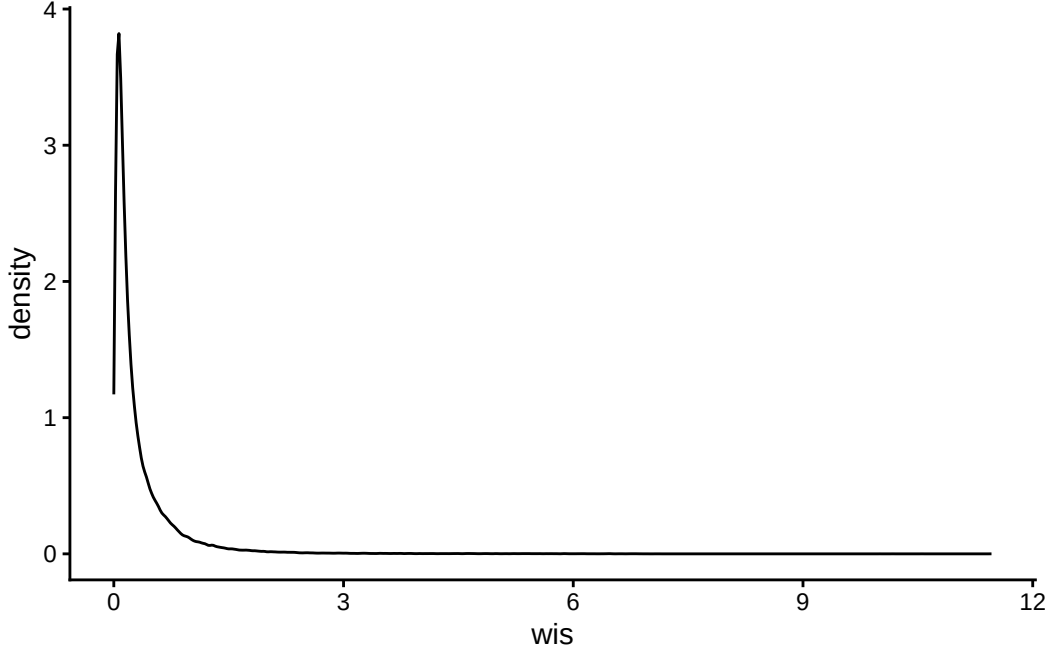


Figure 9: The LWIS is strongly right-skewed, which is what the choice of error family has to accommodate. Density of the LWIS across all scored forecasts.

We therefore compared three error families on this specification, holding the formula and data fixed and varying only the family. The Gaussian family leaves the skew in the residuals rather than modelling it. The Gamma and Tweedie families both accommodate a right-skewed positive outcome, and fit this data almost identically. `mgcv` restricts the Tweedie power parameter to $1 < p < 2$, and the estimate sits near the top of that range, so the data are asking for a variance function close to the Gamma ($p = 2$). The two fit this data almost identically, and both converge. We report the Tweedie fit because a Gamma has no support at zero: fitting it requires adding a constant to every score, which displaces the 553 forecasts that score exactly zero to a value set by the size of that constant.

The choice of family matters a great deal more the log scale than the natural scale. On the log scale, moving from Gaussian to Tweedie reduces residual skew from 5.8 to 0.6 and kurtosis from 77.6 to 10. On the natural scale, the WIS is heavily skewed, and the residual skew is essentially unchanged (1.84); there the change resolves the convergence failure rather than improving the distributional fit.

Table 3: Comparison of error families for the joint model on the log scale. Skew and kurtosis are of the deviance residuals for comparison between family. Lower absolute skew and kurtosis indicate a distributional assumption better matched to the outcome. Tweedie is the family used in the primary model, fitted without the $1e-7$ constant used here (see below).

Family	Deviance explained	AIC	Residual skew	Residual kurtosis
gaussian	0.286	262955	5.85	77.6
Gamma	0.378	-103418	0.52	9.5
Tweedie($p=1.99$)	0.380	-103165	0.58	9.3

A small share of forecasts (553, or 0.27%) score exactly zero: perfect predictions of zero-incidence targets, almost all deaths in the smallest countries. Gamma has no support at zero, so fitting it requires displacing these scores by a small constant. A Tweedie family with power parameter between 1 and 2 has a point mass at zero and admits them directly, so the primary model retains the exact zeros and adds no constant. Adding a constant of $1e-7$ back to every score changes the fit very little (residual skew 0.58 against 0.57, deviance explained 0.380 against 0.382), so the choice is not consequential; we prefer the specification that represents the scores as they are.

We used the `mgcv` package v1.9-4 using R 4.5, with the formula:

```
wis ~ s(Method, Epi_target, bs = "re") + Epi_target + s(CountryTargets, bs = "re") + s(Incidence)
+ s(Trend, bs = "re") + s(Location, bs = "re") + s(VariantPhase, bs = "re") + s(Horizon, k = 4)
+ s(Horizon, Model, k = 4, bs = "fs")
```

We assessed model fit visually plotting observed-versus-fitted values and inspecting residuals in Q-Q and histogram plots.

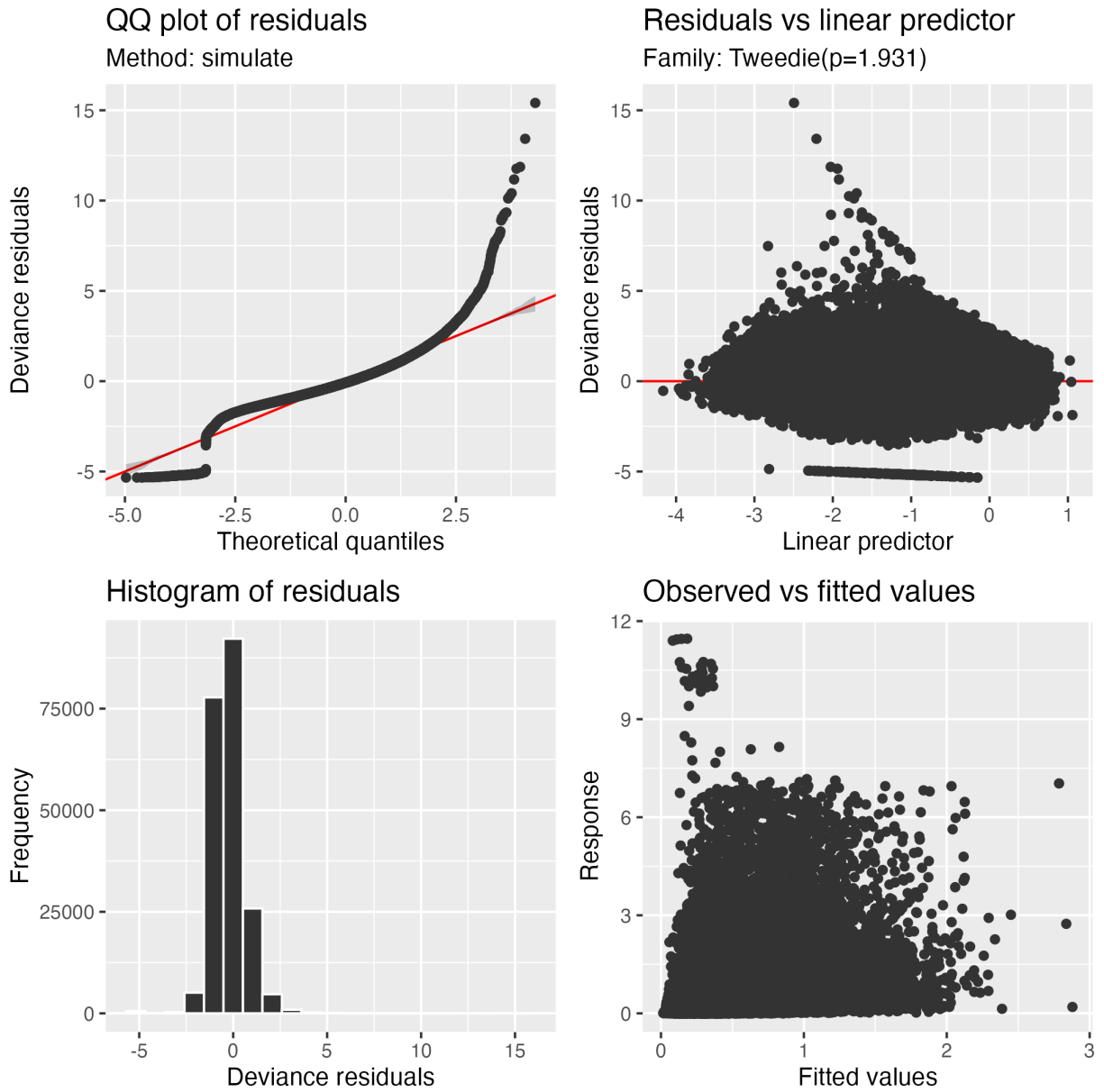


Figure 10: Residual diagnostics for the fitted model: quantile-quantile plot of deviance residuals, residuals against the linear predictor, histogram of residuals, and observed against fitted values.

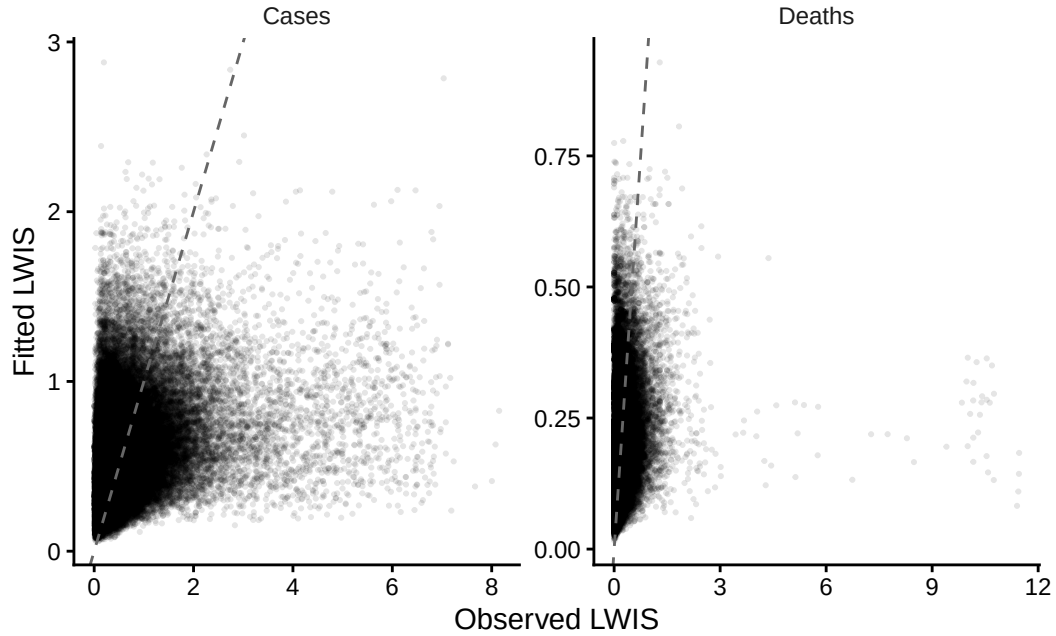


Figure 11: Fitted values track observed scores across the range, with the spread widening at higher scores. Observed against fitted LWIS, by epidemiological outcome.

Estimates

Effects are random effects for each level of a covariate, relative to a forecast at the average level of that covariate with all other terms held fixed. Negative raw effects, or ratios below 1, indicate better-than-average performance.

Partial effects across covariates

Partial effect on forecast performance of key covariates included in the fully adjusted model for additional covariates. Partial effects of individual model structure and individual model are presented in the main text.

(A) Geographic

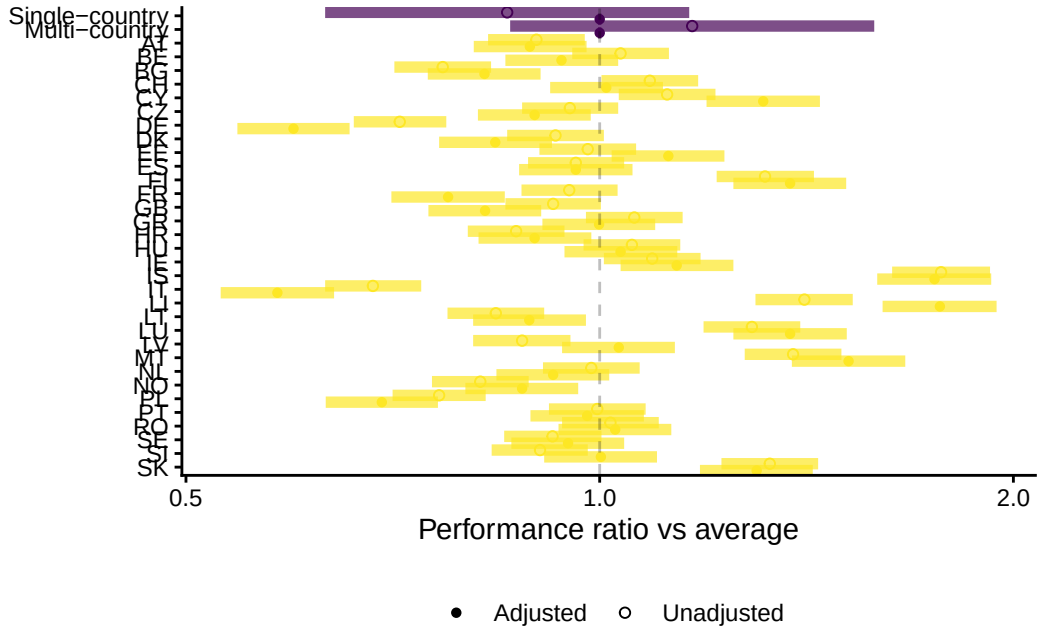


Figure 12: Performance varied more between countries than between single- and multi-country models. Partial effects (95% CI) for geographic scope and country, as a ratio of the expected LWIS relative to the average level.

(B) Epidemic state

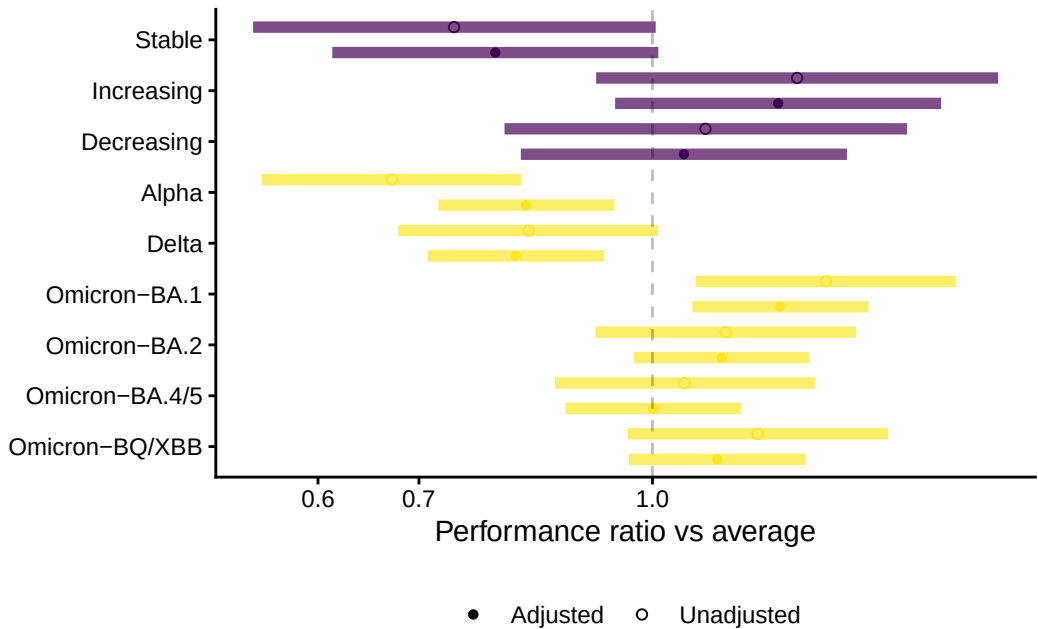


Figure 13: Epidemic trend and variant phase both shifted performance, with increasing trends and the Omicron BA.1 phase hardest to predict. Partial effects (95% CI) as a ratio of the expected LWIS relative to the average level.

Partial effects on the raw scale

The main-text figures and tables report exponentiated effects (multiplicative ratios of the expected LWIS relative to the average level of each factor). The table below gives the underlying raw partial effects from the fully adjusted model.

Table 4: Raw partial effects on the log scale of the linear predictor, from the fully adjusted generalised additive mixed model. Negative values indicate better-than-average performance. 95% CI = 95% confidence interval.

Variable	Group	Partial effect (95% CI), log scale
CountryTargets	Multi-country	0 (-0.003, 0.003)
	Single-country	0 (-0.003, 0.003)
Epi_target	Deaths	-1.042 (-1.165, -0.919)
	AT	-0.117 (-0.211, -0.022)
	BE	-0.064 (-0.158, 0.031)
	BG	-0.193 (-0.288, -0.099)
	CH	0.011 (-0.083, 0.106)
	CY	0.274 (0.179, 0.369)
	CZ	-0.109 (-0.204, -0.015)
	DE	-0.513 (-0.607, -0.419)
	DK	-0.175 (-0.269, -0.08)
	EE	0.115 (0.02, 0.209)
	ES	-0.04 (-0.135, 0.055)
	FI	0.319 (0.224, 0.413)
	FR	-0.254 (-0.349, -0.159)
	GB	-0.192 (-0.287, -0.098)
	GR	-0.001 (-0.096, 0.093)
	HR	-0.109 (-0.203, -0.014)
	HU	0.035 (-0.059, 0.13)
	IE	0.129 (0.035, 0.224)
	IS	0.561 (0.465, 0.656)
	IT	-0.54 (-0.635, -0.445)
	LI	0.57 (0.474, 0.665)
	LT	-0.118 (-0.212, -0.023)
	LU	0.319 (0.224, 0.414)
	LV	0.032 (-0.063, 0.126)

	MT	0.417 (0.322, 0.512)
	NL	-0.078 (-0.173, 0.016)
	NO	-0.13 (-0.225, -0.036)
	PL	-0.365 (-0.459, -0.271)
	PT	-0.021 (-0.116, 0.074)
	RO	0.026 (-0.069, 0.12)
	SE	-0.053 (-0.148, 0.041)
	SI	0.002 (-0.093, 0.096)
	SK	0.263 (0.168, 0.357)
Method	Agent-based	-0.027 (-0.146, 0.091)
	Judgement	-0.036 (-0.152, 0.081)
	Mechanistic	-0.007 (-0.104, 0.09)
	Semi-mechanistic	0.08 (-0.023, 0.183)
	Statistical	-0.01 (-0.11, 0.09)
	AMM-EpiInvert	-0.308 (-0.409, -0.207)
	CovidMetrics-epiBATS	-0.124 (-0.261, 0.014)
	DSMPG-bayes	-0.508 (-0.63, -0.385)
	EuroCOVIDhub-baseline	0.037 (-0.058, 0.132)
	FIAS_FZJ-Epi1Ger	0.263 (0.145, 0.381)
	GoeWroc-BaseBayes	0.483 (0.198, 0.768)
	HZI-AgeExtendedSEIR	-0.157 (-0.268, -0.046)
	ICM-agentModel	0.117 (-0.034, 0.269)
	IEM_Health-CovidProject	-0.027 (-0.113, 0.059)
	ILM-EKF	-0.019 (-0.121, 0.083)
	ITWW-county_repro	0.109 (-0.006, 0.224)
	Imperial-DeCa	0.31 (0.157, 0.463)
	Imperial-RtI0	0.537 (0.385, 0.69)
	Imperial-sbcp	0.45 (0.297, 0.603)
	JBUD-HMXK	0.097 (0.004, 0.189)
	KITmetricslab-bivar_branching	0.012 (-0.326, 0.351)
	Karlen-pypm	-0.155 (-0.243, -0.068)
	LANL-GrowthRate	-0.16 (-0.263, -0.056)
	LeipzigIMISE-SECIR	0.049 (-0.216, 0.314)
	MIMUW-StochSEIR	0.274 (0.108, 0.44)

	MIT_CovidAnalytics-DELPHI	0.143 (0.036, 0.25)
	MOCOS-agent1	-0.396 (-0.546, -0.246)
	MUNI-ARIMA	-0.112 (-0.207, -0.017)
	MUNI-LaggedRegARIMA	-0.181 (-0.299, -0.064)
	MUNI-VAR	0.14 (0.036, 0.243)
	MUNI_DMS-SEIAR	-0.014 (-0.14, 0.112)
	PL_GRedlarski-DistrictsSum	-0.192 (-0.321, -0.062)
	RobertWalraven-ESG	0.092 (-0.003, 0.187)
	SDSC_ISG-TrendModel	0.381 (0.252, 0.511)
	UB-BSLCoV	0.061 (-0.099, 0.221)
	UC3M-EpiGraph	0.313 (0.086, 0.541)
	ULZF-SEIRC19SI	-0.335 (-0.458, -0.213)
	UMass-MechBayes	-0.155 (-0.243, -0.066)
	UMass-SemiMech	-0.061 (-0.167, 0.045)
	UNED-PreCoV2	-0.025 (-0.173, 0.124)
	UNIPV-BayesINGARCHX	0.436 (0.308, 0.564)
	USC-SIkJalpha	0.125 (0.039, 0.21)
	UpgUmibUsi-MultiBayes	-0.391 (-0.584, -0.198)
	bisop-seirfilter	0.004 (-0.216, 0.224)
	bisop-seirfilterlite	0.083 (-0.029, 0.195)
	epiMOX-SUIHTER	-0.348 (-0.506, -0.19)
	epiforecasts-EpiExpert	-0.105 (-0.243, 0.032)
	epiforecasts-EpiExpert_Rt	-0.098 (-0.243, 0.047)
	epiforecasts-EpiExpert_direct	-0.123 (-0.268, 0.023)
	epiforecasts-EpiNow2	-0.025 (-0.127, 0.077)
	epiforecasts-weeklygrowth	-0.2 (-0.298, -0.103)
	itwm-dSEIR	-0.04 (-0.149, 0.069)
	prolix-euclidean	0.525 (0.413, 0.636)
Trend	Decreasing	0.048 (-0.201, 0.297)
	Increasing	0.192 (-0.057, 0.441)
	Stable	-0.24 (-0.489, 0.009)
	Alpha	-0.193 (-0.327, -0.058)
	Delta	-0.209 (-0.343, -0.074)
	Omicron-BA.1	0.195 (0.061, 0.33)

Omicron-BA.2	0.106 (-0.028, 0.24)
Omicron-BA.4/5	0.001 (-0.133, 0.135)
Omicron-BQ/XBB	0.099 (-0.036, 0.234)

The model structure term is fitted as a single grouping factor crossing structure with epidemiological outcome, so its raw partial effects are one per structure-by-outcome cell. The main text reports these exponentiated, and reports a pooled effect per structure derived as a contrast averaging a structure’s two cells.

Table 5: Exponentiated partial effects of model structure by epidemiological outcome, as a ratio of the expected LWIS relative to the average level. Ratios below 1 indicate better-than-average performance.

Model structure	Cases	Deaths
Agent-based	1.09 (0.95, 1.25)	0.87 (0.75, 1)
Judgement	0.95 (0.83, 1.08)	0.98 (0.86, 1.12)
Mechanistic	0.95 (0.84, 1.06)	1.04 (0.93, 1.17)
Semi-mechanistic	1.06 (0.94, 1.2)	1.1 (0.98, 1.24)
Statistical	0.96 (0.85, 1.08)	1.02 (0.91, 1.15)

Table 6: Raw partial effects of model structure by epidemiological outcome, on the log scale, from the fully adjusted generalised additive mixed model. Negative values indicate better-than-average performance. 95% CI = 95% confidence interval.

Model structure	Outcome	Partial effect (95% CI), log scale
Agent-based	Cases	0.088 (-0.05, 0.227)
	Deaths	-0.143 (-0.283, -0.003)
Judgement	Cases	-0.055 (-0.188, 0.079)
	Deaths	-0.017 (-0.15, 0.117)
Mechanistic	Cases	-0.055 (-0.169, 0.06)
	Deaths	0.04 (-0.075, 0.155)
Semi-mechanistic	Cases	0.062 (-0.058, 0.182)
	Deaths	0.099 (-0.021, 0.218)
Statistical	Cases	-0.041 (-0.158, 0.076)
	Deaths	0.02 (-0.097, 0.138)

Model ranking before and after adjustment

Individual models can be ranked by their partial effect on the score, either from a univariate model containing only individual model identity, or from the fully adjusted model. The first ranks models

by observed performance, and so by a combination of the method used and the difficulty of the targets each model chose to forecast. The second ranks them by performance after the target-related covariates are held fixed. The comparison is reported in the main text; the table below gives the models that move furthest between the two.

Table 7: The ten models whose rank changes most under adjustment, of 48 models. A positive change means a model ranks better once the difficulty of its targets is accounted for. Models are anonymised and numbered within model structure.

Model	Unadjusted rank	Adjusted rank	Rank change
Semi-mechanistic 3	4	45	-41
Semi-mechanistic 4	7	48	-41
Semi-mechanistic 2	2	41	-39
Statistical 11	43	7	36
Statistical 10	40	6	34
Mechanistic 15	35	8	27
Semi-mechanistic 10	42	22	20
Statistical 6	32	14	18
Mechanistic 4	12	29	-17
Mechanistic 9	21	38	-17

Individual model effects at one week ahead

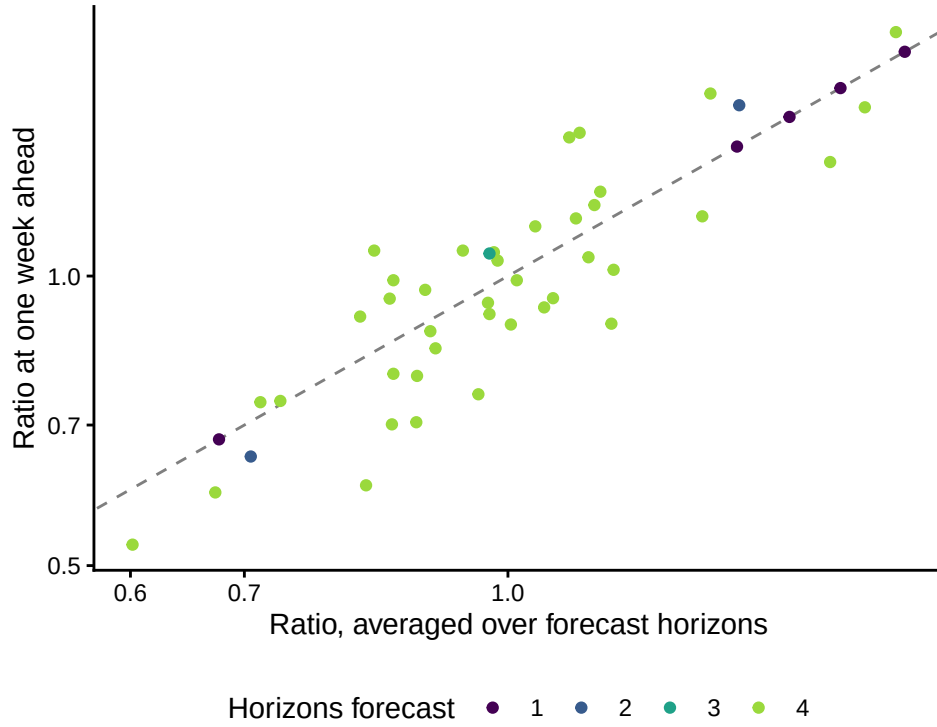


Figure 14: Individual model effects evaluated at one week ahead, the horizon every model forecast, against the effect averaged over the horizons each model forecast (main text). Each point is a model, as the exponentiated deviation from the shared horizon curve; the dashed line marks equality.

Spearman correlation between the two: 0.86.

Sensitivity analyses

Data processing

Population normalisation

We scored forecasts and observations normalised by country population per 100,000. We also scored against the raw total incident weekly count per target. This had near-zero impact on model fitting or results on the log-transformed scores. Difference in median WIS (across all targets):

Table 8: Difference in median WIS between scoring on the raw count and per-100,000 population scales, by epidemiological outcome and scale.

Target	Scale	Raw count	Per 100k
Cases	log	0.292	0.278
Cases	natural	2426	34.4
Deaths	log	0.239	0.093
Deaths	natural	14.7	0.199

Natural scale WIS

We present results from scoring forecast error on the natural scale (difference between observation and prediction in count of case or death incidence), from which the WIS is then calculated and the analysis repeated.

Table 9: Participation and forecasts by model structure and epidemiological outcome, scored on the natural scale.

	Cases	Deaths	Both outcomes	Participation (%)
	Models (%)	Models (%)	Models (%)	
Overall	43 (100%)	39 (100%)	48 (100%)	3%
Method				
Semi-mechanistic	9 (20.9%)	10 (25.6%)	12 (25%)	4%
Statistical	12 (27.9%)	8 (20.5%)	13 (27.1%)	7%
Mechanistic	16 (37.2%)	16 (41%)	17 (35.4%)	3%
Agent-based	3 (7%)	2 (5.1%)	3 (6.2%)	3%
Judgement	3 (7%)	3 (7.7%)	3 (6.2%)	3%

Model specification

We tried several alternative model specifications.

Covariate selection

Table 10: Covariates considered across model versions, the version in which each was introduced, and their impact on model fit and estimates.

Domain	Variable	Model version	Impact
Forecast target	EpiTarget	V2*	None
	VariantPhase	V2*	Some
	Trend	All	Untested
	Horizon	All	Untested

Domain	Variable	Model version	Impact
Forecaster strategy	Incidence	All	Untested
	Location	All	Untested
	CountryTargets	All	Untested
	Model	All	Untested
Outcome	Method	All	

* Covariates included in V2 were added based on reviewer feedback.

Error family and link

We first used a Gaussian distribution with a log link. Because LWIS is strongly right-skewed, this left high residual skewness (5.8) and kurtosis (77.6). We avoid using additional data transformations (e.g. an additional log transform) on the outcome (LWIS) as this would violate propriety of the score. We instead addressed the skew through the error family, comparing Gaussian, Gamma and Tweedie families as reported under [Model fitting](#) above, and adopting the Tweedie family for the reported analysis.

The fitted relationships and partial effects remained stable across parameterisations.

Structure by outcome under a Gaussian family

The structure-by-outcome effects are sensitive to the error family, in magnitude throughout and in direction for two of the five structures. The table below refits the same specification with a Gaussian family and log link, alongside the primary Tweedie fit. The case-versus-death contrast keeps its sign under the Gaussian family for judgement, mechanistic and statistical models. It does not for agent-based models, where the separation vanishes, or for semi-mechanistic models, where it reverses; in both the Gaussian contrast is close to zero. The Gaussian fit is dominated by the largest scores, which fall disproportionately among case forecasts, so it weights precisely the observations the Tweedie family is chosen to accommodate. Including or excluding the Hub baseline model makes no material difference to these estimates. We report the Tweedie fit throughout; the size of any single structure-by-outcome contrast should be read with this sensitivity in mind.

Table 11: Performance ratio by model structure and epidemiological outcome, under the primary Tweedie family and under a Gaussian family with a log link. A ratio below 1 indicates better-than-average performance.

Model structure	Outcome	Tweedie	Gaussian
Agent-based	Cases	1.09	0.99
	Deaths	0.87	1.00
Judgement	Cases	0.95	0.97
	Deaths	0.98	1.03
	Cases	0.95	0.98

Mechanistic	Deaths	1.04	1.00
	Cases	1.06	1.06
Semi-mechanistic	Deaths	1.10	0.97
	Cases	0.96	1.00
Statistical	Deaths	1.02	1.01
	Cases	0.96	1.00