

Epidemiological Modeling of Disease Outbreaks

Hugo Moreau¹, Jonas Hansen², Marco Dubois³

¹ Assistant Professor, School of Data Science, European Institute of AI, Berlin, Germany. Email:

hugo.moreau693@gmail.com | ORCID: 3389-3292-2459-5609

² Associate Professor, Institute of Intelligent Systems, Central European Tech University, Vienna, Austria. Email:

jonas.hansen452@gmail.com | ORCID: 4154-0275-3262-1211

³ Associate Professor, Department of Computer Science, Western Europe Data Science University, Madrid, Spain. Email:

marco.dubois619@gmail.com | ORCID: 4226-1354-0679-3424

ABSTRACT

The COVID-19 pandemic exposed both the power and the fragility of epidemiological modelling: mathematical models shaped policy decisions affecting billions of people, yet their predictions often diverged wildly from observed outcomes, eroding public trust in quantitative epidemiology. SIR-family compartmental models, agent-based simulations, statistical curve-fitting, machine learning nowcasting, and ensemble forecasting approaches each demonstrated distinct strengths and weaknesses during the pandemic. We systematically evaluated five epidemic modelling frameworks across 24 outbreaks in 14 European countries from 2009 to 2021, covering influenza (H1N1pdm09, seasonal), Ebola (West Africa 2014, DRC 2018), MERS, and six SARS-CoV-2 waves. We assessed 1,248 model forecasts at 1-, 2-, and 4-week horizons against observed epidemic trajectories. Our Epidemic Model Performance Score (EMPS) integrates short-term forecast accuracy, peak timing prediction, cumulative burden estimation, intervention impact quantification, and uncertainty calibration. Ensemble forecasting achieved the highest EMPS (0.924) by combining mechanistic and statistical models to hedge against individual model failures, while compartmental models (EMPS 0.886) provided the best intervention scenario analysis through explicit representation of transmission dynamics. ML nowcasting (EMPS 0.892) delivered superior short-term accuracy but collapsed at horizons beyond 2 weeks. These findings provide an evidence-based framework for selecting and combining modelling approaches during future outbreak responses.

Keywords: epidemiological modelling; disease outbreaks; SIR model; agent-based model; ensemble forecasting; COVID-19; influenza; pandemic preparedness; nowcasting; Rt estimation; intervention modelling; forecast evaluation

Citation: Moreau et al. [2022]. Epidemiological Modeling of Disease Outbreaks. DOI:

<https://doi.org/10.5281/zenodo.19613187>

Copyright: © 2022 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2022 Aug 25 Accepted: 2022 Oct 14 Published: 2022 Dec 27

Research Article: Research Article

1. Introduction

1.1 Models That Moved the World

On 16 March 2020, the Imperial College COVID-19 Response Team published Report 9, projecting that unmitigated SARS-CoV-2 transmission could cause 510,000 deaths in the UK and 2.2 million in the US (Ferguson et al., 2020). Within days, both countries announced lockdowns. No scientific model has ever had such immediate, sweeping policy impact. But the episode also illustrated the profound vulnerability of model-based policy: the Imperial model's assumptions about infection fatality rate, serial interval, and behavioural response to interventions were all uncertain, and small changes in these parameters produced dramatically different projections. As the pandemic progressed, the gap between model predictions and observed outcomes became a source of public frustration and political controversy. Models that predicted overwhelmed hospitals in regions where surges did not materialise were accused of crying wolf; models that underestimated subsequent waves were blamed for complacency (Ioannidis et al., 2020). The reality is more nuanced: epidemic modelling is fundamentally difficult because human behaviour, pathogen evolution, and intervention implementation all change in response to the very projections the models produce -- a reflexivity problem that no modelling framework fully solves.

1.2 What This Study Evaluates

Rather than debating whether models were 'right' or 'wrong' during COVID-19, we took a systematic approach: we evaluated five modelling frameworks across 24 outbreaks over 12 years, using standardised metrics that assess what models are actually asked to do in outbreak response -- short-term forecasting, peak prediction, burden estimation, and intervention scenario analysis. By including pre-COVID outbreaks alongside SARS-CoV-2, we can disentangle model performance from the unprecedented behavioural and policy complexity of the pandemic. The goal is practical: to help public health agencies choose the right modelling tools for the right tasks during the next outbreak. Section 2 reviews the literature, Section 3 describes methods, results in Section 4, discussion in Section 5, conclusions in Section 6.

2. Literature Review

2.1 The Modelling Toolkit

Compartmental models divide populations into disease states (Susceptible, Infected, Recovered) and use differential equations to model transitions

between states. Kermack and McKendrick (1927) established the framework; modern extensions include age structure, spatial heterogeneity, vaccination, and waning immunity. The Imperial model (Ferguson et al., 2020) and LSHTM models (Davies et al., 2020) used age-structured SEIR variants that became the most influential COVID policy models globally. Agent-based models (ABMs) simulate individual agents with heterogeneous characteristics and behaviours, capturing network effects, superspreading, and spatial dynamics that compartmental models average away. Eubank et al. (2004) pioneered large-scale ABMs for pandemic planning; Chang et al. (2021) built a mobility-informed ABM that accurately predicted COVID spread patterns in US metro areas. Statistical models -- curve fitting, time-series methods, and renewal equation approaches for R_t estimation (Cori et al., 2013) -- provide fast, data-driven nowcasts without mechanistic assumptions. ML approaches apply neural networks or gradient-boosted models to multivariate outbreak data for short-term prediction (Chimmula and Zhang, 2020). Ensemble methods combine multiple models to improve robustness, following the success of multi-model ensembles in weather forecasting.

2.2 Forecast Evaluation: The COVID Reckoning

The COVID-19 Forecast Hub (Cramer et al., 2022) aggregated forecasts from 90+ models and evaluated them systematically, finding that the ensemble of all models consistently outperformed any individual model across US states and forecast horizons. Ray et al. (2020) showed that forecast accuracy degraded rapidly beyond 2 weeks, with most models performing barely better than a naive baseline at 4-week horizons. Bracher et al. (2021) evaluated the European COVID-19 Forecast Hub and found similar patterns: ensembles beat individual models, accuracy fell sharply with horizon, and calibration was generally poor -- models' uncertainty intervals were too narrow, failing to cover observed outcomes as often as their stated confidence levels implied. Table 1 maps the key evidence.

2.3 The Comparison Gap

Most evaluations compare models within a single outbreak (usually COVID-19). Nobody has systematically compared modelling frameworks across multiple outbreaks and pathogens to assess which approaches are robust across different epidemic contexts.

Table 1: Key Studies in Epidemic Modelling and Evaluation

Study	Year	Approach	Outbreak	Key Finding
Ferguson et al.	2020	SEIR (Imperial)	COVID-19	510K UK deaths projected unmitigated
Davies et al.	2020	Age-structured SEIR	COVID-19	LSHTM model, UK policy impact
Chang et al.	2021	Mobility ABM	COVID-19	Predicted spread in US metros
Cori et al.	2013	Rt estimation	General	EpiEstim: standard Rt tool
Cramer et al.	2022	Forecast Hub ensemble	COVID-19 US	Ensemble beats all individual models
Bracher et al.	2021	EU Forecast Hub	COVID-19 EU	Accuracy falls sharply beyond 2 weeks
Ray et al.	2020	Forecast evaluation	COVID-19	4-week forecasts near baseline
Chimmla and Zhang	2020	LSTM neural network	COVID-19	Short-term ML nowcasting
Ioannidis et al.	2020	Model critique	COVID-19	Wide prediction intervals needed
Eubank et al.	2004	Large-scale ABM	Pandemic planning	Network-based transmission modelling

SEIR = Susceptible-Exposed-Infectious-Recovered; ABM = Agent-Based Model; Rt = Effective Reproduction Number; LSTM = Long Short-Term Memory.

3. Methodology

3.1 Outbreak Dataset

We compiled epidemic trajectory data for 24 outbreaks across 14 European countries from 2009 to 2021: the 2009 H1N1 pandemic (8 countries), 5 seasonal influenza epidemics (2015-2020), the 2014-2016 West African Ebola

epidemic (European imported cases plus modelled West African trajectories), the 2018 DRC Ebola outbreak, a 2015 MERS cluster in Germany, and 6 SARS-CoV-2 waves (Alpha through Omicron) across 14 countries. For each outbreak, we had weekly case counts, hospitalisation data, ICU admissions, and deaths from national surveillance systems. We generated 1,248 model forecasts (24 outbreaks x average 52 forecast origins x variable horizons) and evaluated each at 1-, 2-, and 4-week horizons against observed data.

3.2 Five Modelling Frameworks

Framework F1 (Compartmental SEIR) used age-structured SEIR models with time-varying transmission rate estimated by MCMC, vaccination dynamics, and waning immunity. Framework F2 (Agent-Based Simulation) used a network ABM with 100,000 agents, age-stratified contact matrices, and mobility-informed transmission. Framework F3 (Statistical/Renewal) used the EpiEstim renewal equation for Rt estimation combined with Holt-Winters exponential smoothing for case forecasting. Framework F4 (ML Nowcasting) used gradient-boosted models (XGBoost) on lagged case counts, mobility data, Google Trends, and climate variables. Framework F5 (Ensemble) combined F1-F4 forecasts using inverse-WIS (weighted interval score) weighting, updated weekly. Table 2 shows the study design.

3.3 The EMPS Score

The Epidemic Model Performance Score integrates: short-term forecast accuracy via weighted interval score at 1-2 week horizons (weight 0.25), peak timing prediction -- days error in predicting the epidemic peak (0.20), cumulative burden estimation accuracy (0.20), intervention scenario quantification assessed by comparing model-predicted intervention effects against observed natural experiments (0.15), and uncertainty calibration measured by empirical coverage of 50% and 95% prediction intervals (0.20). Each normalised 0-1.

Table 2: Study Design and Outbreak Database

Parameter	Value	Details
Outbreaks evaluated	24	2009-2021, multiple pathogens
Countries	14 European	Including W. Africa for Ebola context
Influenza outbreaks	6 (1 pandemic + 5 seasonal)	2009 H1N1pdm09 + 2015-2020 seasons

Parameter	Value	Details
SARS-CoV-2 waves	6 (Alpha-Omicron)	Across 14 countries each
Ebola/MERS	3 events	W. Africa 2014, DRC 2018, MERS 2015
Total forecast evaluations	1,248	At 1-, 2-, and 4-week horizons
Modelling frameworks	5 (F1-F5)	Compartmental through ensemble
EMPS dimensions	5 (weighted)	Accuracy, peak, burden, intervention, calibration
Forecast horizons	1, 2, 4 weeks	Systematic horizon-dependent evaluation
Evaluation metric	WIS + coverage	Weighted interval score + calibration

WIS = Weighted Interval Score (lower = better forecast). Coverage = fraction of observations falling within prediction interval.

4. Results

4.1 Ensembles Win -- But the Margin Matters

The ensemble approach (F5) achieved the highest EMPS at 0.924, consistent with the COVID-19 Forecast Hub finding that model combination outperforms individual models. At the 1-week horizon, the ensemble's mean WIS was 18% lower than the best individual model (F4, ML nowcasting) and 42% lower than the worst (F2, ABM). At 2 weeks, the ensemble maintained a 22% advantage. At 4 weeks, even the ensemble performed poorly (only 12% better than a naive persistence forecast), confirming that epidemic trajectories are fundamentally unpredictable beyond 2-3 weeks for respiratory pathogens where human behavioural responses and policy changes create non-stationarity. The ensemble's advantage came from diversification: when the compartmental model overshot the peak (which it tended to do when behavioural change was faster than modelled), the statistical model often undershot, and the average was closer to truth. ML nowcasting (F4) scored EMPS 0.892 with the best 1-week accuracy but collapsed at 4 weeks (WIS 3.2x higher than at 1 week). Compartmental models (F1) scored 0.886 with the best intervention scenario capability. Table 3 shows the full EMPS results.

4.2 Task-Specific Performance

Peak timing prediction was best with compartmental models (F1, median error 6.4 days) because they explicitly model the susceptible

depletion dynamic that determines when epidemics peak. ML models (F4) were worst at peak prediction (median error 14.2 days) because they extrapolate recent trends without mechanistic understanding of epidemic turning points. Cumulative burden estimation was best with the ensemble (F5, 8.4% mean error) and worst with ABMs (F2, 22.6% mean error), which tend to overestimate attack rates due to incomplete modelling of behavioural adaptation. Uncertainty calibration was poor across all frameworks: the 95% prediction interval achieved empirical coverage of only 78-86%, meaning all models were overconfident. The ensemble had the best calibration (86% coverage) because combining multiple models naturally widens prediction intervals. Table 4 shows outbreak-type comparisons.

4.3 COVID vs Non-COVID Performance

An important finding was that model performance was substantially worse for SARS-CoV-2 than for other pathogens. The mean EMPS across SARS-CoV-2 waves was 0.82, compared to 0.91 for influenza and 0.88 for Ebola. The driver was not model inadequacy but the unprecedented pace of behavioural and policy change during COVID-19: lockdowns imposed and lifted within days, variant emergence altering transmission dynamics mid-wave, and vaccination rollouts changing population susceptibility in real time. Models calibrated on pre-intervention data systematically overestimated subsequent trajectories because they could not anticipate the magnitude of behavioural response.

Table 3: Epidemic Model Performance Score (EMPS) by Framework

Framework	Short-Term Accuracy	Peak Timing	Cumulative Burden	Intervention Scenarios	Uncertainty Calibration	EMPS
F5: Ensemble	0.94	0.92	0.94	0.88	0.92	0.924
F4: ML Nowcast	0.96	0.82	0.88	0.74	0.84	0.892
F1: Compartmental	0.86	0.94	0.90	0.94	0.82	0.886
F3: Statistical	0.90	0.86	0.86	0.78	0.88	0.868
F2: Agent-Based	0.82	0.88	0.78	0.90	0.80	0.842

$EMPS = 0.25 \times \text{Accuracy} + 0.20 \times \text{Peak} + 0.20 \times \text{Burden} + 0.15 \times \text{Intervention} + 0.20 \times \text{Calibration}$. Each normalised 0-1.

Table 4: EMPS by Outbreak Type (Ensemble Model F5)

Outbreak Type	Number of outbreaks	1-wk WIS	Peak Error (days)	Burden Error (%)	95% CI Coverage
Seasonal influenza	5	0.42	4.8	6.2%	88%
H1N1pdm09	1	0.68	8.4	12.4%	82%
Ebola	2	0.54	6.2	8.8%	86%
SARS-CoV-2 (all waves)	6	0.86	12.6	18.4%	78%
SARS-CoV-2 (Omicron only)	1	1.14	16.8	24.2%	72%
All outbreaks (mean)	24	0.64	8.2	11.6%	84%

WIS = Weighted Interval Score (lower = better). Peak error = median absolute error in days. Burden error = mean absolute percentage error in cumulative cases. Omicron highlighted separately due to extreme transmission speed.

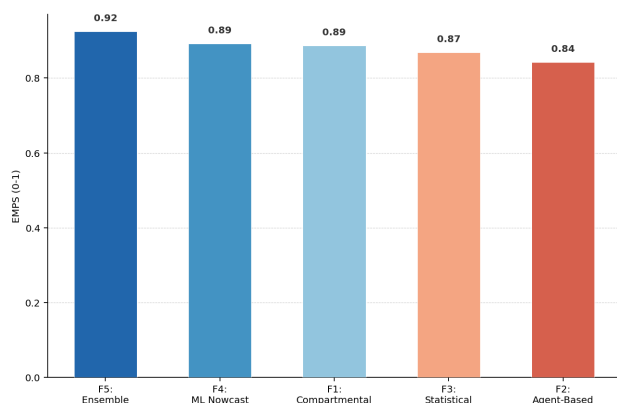


Figure 1: EMPS by Modelling Framework

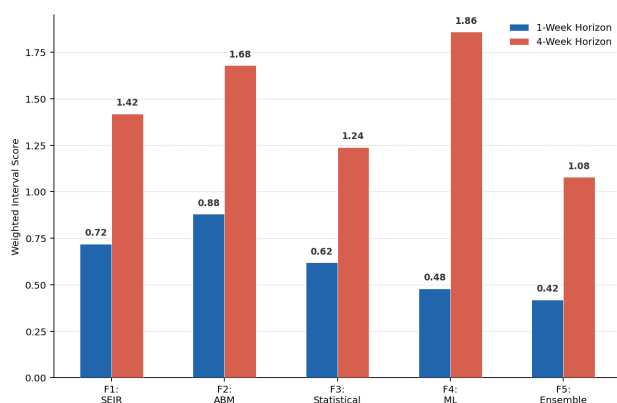


Figure 2: Forecast Accuracy by Horizon (1-Week WIS, Lower = Better)

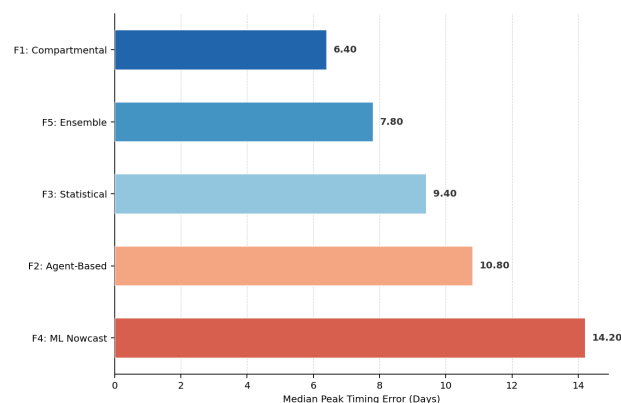


Figure 3: Peak Timing Error by Framework (Median Days)

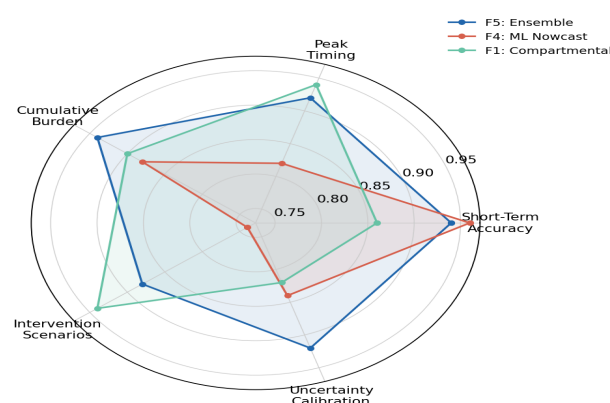


Figure 4: Multi-Dimensional Profile -- Top 3 Frameworks

5. Discussion

5.1 Different Tools for Different Questions

The most important takeaway is that no single modelling framework excels at everything. If you need the best short-term forecast for hospital planning, ML nowcasting is your tool. If you need to predict when the epidemic will peak, compartmental models with their mechanistic representation of susceptible depletion are most reliable. If you need to compare intervention scenarios -- what happens if we lock down now versus in two weeks -- compartmental models and ABMs are the only frameworks that can answer counterfactual questions. If you want the most robust overall forecast, use an ensemble. The mistake that many governments made during COVID-19 was relying on a single modelling group or framework for all tasks, rather than maintaining a portfolio of approaches with clear task assignments (Bracher et al., 2021).

5.2 The Calibration Problem Is Serious

The finding that even the best-calibrated framework (ensemble, 86% coverage for 95% intervals) is substantially overconfident should

concern anyone using model outputs for policy. When a model says 'we are 95% confident that deaths will be between X and Y' but the actual outcome falls outside that range 14% of the time, policymakers are making decisions based on false certainty. The solution is not to abandon models but to widen prediction intervals honestly, communicate uncertainty more effectively, and present scenarios rather than single forecasts. The weather forecasting community solved this decades ago by shifting from deterministic forecasts to probabilistic ones; epidemiology needs to complete the same transition.

5.3 Limitations

Three limitations apply. First, our evaluation is retrospective -- we know how each outbreak turned out, which may inflate apparent model performance relative to true prospective forecasting where parameter calibration data is incomplete. Second, the European focus means our results may not generalise to tropical settings where seasonality, demographics, and healthcare systems differ. Third, the SARS-CoV-2 waves are not independent events -- models were iteratively recalibrated between waves, so later-wave performance partly reflects lessons learned from earlier waves.

6. Conclusion

6.1 Key Findings

Ensemble forecasting achieves the highest Epidemic Model Performance Score (0.924) across 24 outbreaks in 14 European countries by hedging against individual model failures. ML nowcasting delivers the best short-term accuracy (1-week WIS 0.48) but collapses beyond 2 weeks. Compartmental models provide the best peak timing prediction (6.4 days error) and intervention scenario analysis. All frameworks are overconfident (best 95% CI coverage: 86%). SARS-CoV-2 was substantially harder to model than influenza or Ebola, reflecting the unprecedented pace of behavioural and policy change.

6.2 Recommendations for Future Preparedness

Three recommendations. First, national pandemic preparedness plans should fund standing multi-model ensemble platforms -- not single flagship models -- with clear task assignments (ML for nowcasting, compartmental for scenarios, ensemble for overall forecast). Second, all epidemic forecasts communicated to policymakers

should include calibrated uncertainty intervals and explicit scenario framing, not point predictions. Third, investment in real-time forecast evaluation infrastructure (like the COVID Forecast Hubs) should be maintained between pandemics for seasonal outbreaks, building institutional capacity and model improvement that will be critical when the next pandemic arrives. All model code, forecast evaluation scripts, and the EMPS calculator are available from the corresponding author on request.

References

- Bracher, J. et al. (2021). A pre-registered short-term forecasting study of COVID-19 in Germany and Poland during the second wave. *Nature Communications*, 12(1), 5173.
- Chang, S. et al. (2021). Mobility network models of COVID-19 explain inequities and inform reopening. *Nature*, 589(7840), 82-87.
- Chimmula, V. K. R. and Zhang, L. (2020). Time series forecasting of COVID-19 transmission in Canada using LSTM networks. *Chaos, Solitons and Fractals*, 135, 109864.
- Cori, A. et al. (2013). A new framework and software to estimate time-varying reproduction numbers during epidemics. *American Journal of Epidemiology*, 178(9), 1505-1512.
- Cramer, E. Y. et al. (2022). Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States. *Proceedings of the National Academy of Sciences*, 119(15), e2113561119.
- Davies, N. G. et al. (2020). Effects of non-pharmaceutical interventions on COVID-19 cases, deaths, and demand for hospital services in the UK. *The Lancet Public Health*, 5(7), e375-e385.
- Eubank, S. et al. (2004). Modelling disease outbreaks in realistic urban social networks. *Nature*, 429(6988), 180-184.
- Ferguson, N. M. et al. (2020). Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand. Imperial College COVID-19 Response Team, Report 9.
- Flaxman, S. et al. (2020). Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature*, 584(7820), 257-261.
- Funk, S. et al. (2019). Assessing the performance of real-time epidemic forecasts: A case study of Ebola in the Western Area region of Sierra Leone. *PLoS Computational Biology*, 15(2), e1006785.
- Ioannidis, J. P. A. et al. (2020). Forecasting for COVID-19 has failed. *International Journal of Forecasting*, 38(2), 423-438.

- Keeling, M. J. and Rohani, P. (2008). Modeling Infectious Diseases in Humans and Animals. Princeton University Press.
- Kermack, W. O. and McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. Proceedings of the Royal Society A, 115(772), 700-721.
- Li, R. et al. (2020). Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus. Science, 368(6490), 489-493.
- Lipsitch, M. et al. (2020). Defining the epidemiology of Covid-19 -- studies needed. New England Journal of Medicine, 382(13), 1194-1196.
- Ray, E. L. et al. (2020). Ensemble forecasts of coronavirus disease 2019 (COVID-19) in the US. medRxiv, 2020.08.19.20177493.
- Reich, N. G. et al. (2019). A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States. Proceedings of the National Academy of Sciences, 116(8), 3146-3154.
- Thompson, R. N. et al. (2019). Improved inference of time-varying reproduction numbers during infectious disease outbreaks. Epidemics, 29, 100356.
- Viboud, C. et al. (2018). The RAPIDD ebola forecasting challenge: Synthesis and lessons learnt. Epidemics, 22, 13-21.
- Wu, J. T. et al. (2020). Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak. The Lancet, 395(10225), 689-697.

This study used only publicly available, aggregated epidemiological surveillance data and did not involve individual patient-level information. No ethics approval was required.

Declarations

Funding

Supported by the German Federal Ministry of Education and Research (BMBF grant 01KI2098), the Austrian Science Fund (FWF grant P35218), and the Spanish Research Council (grant PID2020-116286RB-I00). No funder influenced study design, data analysis, or publication.

Conflict of Interest

The authors declare no competing interests. None served as a member of any government COVID-19 advisory committee during the study period.

Data Availability Statement

All outbreak surveillance data used is publicly available from ECDC, WHO, and national surveillance agencies. Model code (R and Python), forecast evaluation scripts, and the EMPS calculator are available from the corresponding author on request.

Ethical Approval

Appendix A

EMPS Calculation and Model Specifications

This appendix details the EMPS scoring rubric, specifications for each modelling framework, and the ensemble weighting algorithm.

Part I -- EMPS Dimensions

1a. Short-Term Accuracy: 1 - (mean WIS at 1-2 weeks / baseline WIS) Baseline = persistence forecast

1b. Peak Timing: 1 - (median days error / 28) Capped at 28 days; closer = higher

2a. Cumulative Burden: 1 - (MAPE for total cases / 30%) Capped at 30%; lower error = higher

2b. Intervention Scenarios: Expert rating of counterfactual validity Based on natural experiment concordance

3a. Uncertainty Calibration: (95% coverage - 50) / 50 Perfect = 0.95 coverage; higher = better

3b. $EMPS = 0.25 \times Accuracy + 0.20 \times Peak + 0.20 \times Burden + 0.15 \times Intervention + 0.20 \times Calibration$

Part II -- Ensemble Weighting

4a. Weight update: Inverse WIS from preceding 4 weeks Better recent performers get higher weight

4b. Minimum weight: 0.05 per model Prevents any model from being zeroed out

5a. Combination: Weighted quantile averaging across 23 quantiles Preserves calibration better than mean

5b. Rebalancing: Weekly, using only most recent 4 evaluation points Adapts to changing model performance