Multi-model ensembles for infectious disease forecasting: A protocol for a systematic review

Background

Rationale

Infectious disease modelling is a useful tool for supporting outbreak control, offering to interpret the complex uncertainty of epidemiological dynamics. Modellers handle this uncertainty with a variety of approaches, choices, and interpretations during modelling work. Working in collaboration offers comparability across this diversity of modelling work. Modelling collaborations may aim to enable expert elicitation among modellers, clarify the extent and policy relevance of uncertainty, or provide a synthesis of modelling evidence (1–4). Specifically, collaboration among multiple independent and diverse modelling teams may create a stronger basis for evidence-informed policy support (5,6).

Outputs from such modelling collaborations often include a quantitative combination of numerical model results into an ensemble projection. A key benefit of model combination is the increased predictive accuracy of the combined result (7,8). Such findings underlie ensemble approaches in, for example, economics, logistics, or meteorology (9). This better performance partly comes from the independent information added by each projection, while reducing the variance in uncertainty across projections. Meanwhile, the unreliability of individual model performance may mean that there is no obvious best way to combine projections beyond a simple linear average (the "forecast combination puzzle" (10)). These results appear to hold in infectious disease settings, even with an increasing number and range of approaches and methods for model combination (e.g. (11–13)).

This review aims to summarise existing evidence on the predictive accuracy of multi-model ensemble projections of infectious disease (forecasts). As a secondary aim, this review will capture some of the benefits and challenges involved in such collaborations among modellers. This will draw together an increasing literature analysing multi-model collaborations, and support their future design, communication, and evaluation.

Objectives

To assess the accuracy and value of multi-model ensembles for forecasting infectious disease outbreaks.

RQ1: What is the predictive performance of multi-model ensemble projections from independent models in comparison to individual component models when forecasting infectious disease?
RQ2: What are the benefits and challenges of such multi-model ensembles?

Review framework

Component	Description	Inclusion criteria	Exclusion criteria
Population	Population-level projections of infectious disease characteristics	Human populations Any infectious disease outbreak (e.g., influenza, COVID-19, MERS, Ebola, dengue); may include seasonal epidemics and pandemics Any infectious disease characteristic (e.g., incidence, peak timing, final size)	Animal populations Non-communicable diseases Individual patient level predictions e.g., for clinical risk
Intervention	Multi-model combination	Projections combining ≥2 models from independent research groups Any combination method (e.g., linear averaging, weighted averaging, Bayesian model averaging) Qualitative combination methods (e.g. expert elicitation, Delphi method)	Studies of single models only Multiple models compared without combination
Comparator	Component models from independent research groups	Single projections from component models within the ensemble	Studies not reporting component models All component models created by a single research group
Outcomes	**Primary Outcomes:** Relative performance evaluation of predictive accuracy **Secondary Outcomes:** Evaluation of overall performance against projection target(s)	**Primary:** Absolute or relative measures of predictive performance (e.g., differences in mean absolute error, coverage of prediction intervals, interval scores, relative skill scores, bias) **Secondary:**	Studies focusing solely on model methodology without performance evaluation Evaluation only of either ensemble or single models, without comparison

	Benefits and challenges of ensemble approaches	Evaluation of predictive performance given characteristics of the target time-series, e.g. epidemic phase, overall predictability Evaluations of process for modelling collaboration/combination (e.g., modeller onboarding, resource requirements, time to consensus) Evaluation of impacts of modelling collaboration/combination (e.g. policy relevance, communication effectiveness, stakeholder acceptance)	
Study Design	Empirical studies evaluating model projections	Retrospective, real-time, or prospective analyses Case studies of outbreak forecasting exercises Mixed-methods studies including qualitative evaluation	Literature reviews, letters, commentaries, and editorials

Search strategy

Sources

We will search published literature in MEDLINE and EMBASE databases, and preprint servers medRxiv, bioRxiv, and aRxiv, to capture reporting from real-time outbreak work (14).

Search terms

We developed search terms using a combination of expert judgement, a test set of relevant papers, and generative AI to refine search terms. We followed a similar search strategy to Pollett et al (15), combining three sets of terms (infectious disease, forecasting, and our focus on model combination). We validated this against a benchmark of twenty articles indexed in Pubmed using SRAccelerator software (16), with 100% recall. We then used the generative AI (Claude, Anthropic) to suggest synonyms or redundancies. KS developed the search terms, reviewed by SF, and we ran the final search on 6 June 2025.

We accessed medRxiv and bioRxiv records using the medrxivr R interface to the API. We will de-duplicate records using Rayyan or EPPI-Reviewer. We will consider using the ASReview software (17) to prioritise record screening. After full text screening, we will use forward and backward citation searching to further identify missing records.

Source	Search terms	Records
MEDLINE	(("infectious disease".ti,ab. OR epidemic$.ti,ab. OR pandemic$.ti,ab. OR outbreak$.ti,ab. OR influenza.ti,ab. OR COVID$.ti,ab. OR virus.ti,ab.) AND (nowcast$.ti,ab. OR forecast$.ti,ab. OR predicted.ti,ab. OR prediction$.ti,ab. OR predictive.ti,ab. OR projected.ti,ab. OR projection$.ti,ab.) AND ("ensemble".ti,ab. OR "multi-model$".ti,ab. OR "multi model$".ti,ab. OR "multiple model$".ti,ab. OR stacking.ti,ab. OR "model averaging".ti,ab. OR "forecast combination".ti,ab. OR "model combination".ti,ab.))	1007
Embase	As MEDLINE	1354
medRxiv/bioRxiv	("infectious disease" OR "epidemic*" OR "pandemic*" OR "outbreak*" OR "influenza" OR "COVID*" OR "virus") AND ("nowcast*" OR "forecast*" OR "predicted*", "prediction*", "predictive", "projected", "projection*") AND ("ensemble" OR "multi-model*" OR "multi model*" OR "multiple model*" OR "stacking" OR "model averaging" OR "forecast combination" OR "model combination")'	48
aRxiv	As medRxiv	185

Screening

We will import records into either Rayyan or EPPI-reviewer for deduplication, using automated deduplication with human supervision. One reviewer will screen on titles and abstract, blind to authors, journal, and other metadata. We are unable to use multiple reviewers at this stage due to resource limitations. Two reviewers will screen the remaining articles using full text.

Data extraction

We aim to extract the following from each study where possible. Key data are identified in **bold**.

Component	Data
Population	*Study* - Authors - Year - DOI - Data and code availability *Setting* - Pathogen/disease - Time period - Epidemiological target *Data collection* - Aims and context of the work - Projection horizon - Projection data collection method
Intervention	*Multi-model combination* - Model selection or inclusion criteria - Combination method(s) used - Combination performed prospectively or retrospectively - Number and variation in model components
Comparator	*Component models from independent research groups* - Recruitment method for modelling teams - Summary characteristics of contributing teams and models - Projections contributed prospectively or retrospectively

Outcomes	*Primary outcomes:* - **Comparative performance of combined projection against component models** - Characteristics of ensemble performance against observed data - Evaluation of overall predictability of the projection target *Secondary outcomes:* - Evaluation of process (e.g., modeller onboarding, resource requirements, time to consensus) - Evaluation of impacts (e.g. policy relevance, communication effectiveness, stakeholder acceptance) - Other benefits/advantages of multi-model ensemble - Other challenges/limitations of multi-model ensemble
Study Design	*Empirical studies evaluating model projections* - Performance evaluation metric(s) - Evaluation method pre-specified or post-hoc - Inclusion criteria for evaluation

Quality and risk of bias

We use items reported against the EPIFORGE checklist to assess study quality. We consider confounding, selection, information, and reporting biases, using the ROBINS-I tool for assessing risk of bias in non-randomised intervention studies.

Analysis

We will produce a narrative synthesis of reported accuracy of the ensemble forecast against observed data in absolute terms, and if available, relative to the accuracy of individual model components.

Where possible, quantitative synthesis methods will aim to include:

- Meta-analysis of comparable evaluation metrics, e.g. absolute error
- Summarising effect estimates (magnitude and range of effects)
- Vote-counting direction of effect

Potential sub-group analyses, where available, may include:

- Ensemble method
- Characteristics of contributing models
- Forecast horizon
- Outbreak

Code and data availability

We publish all code and data associated with this work at:
https://github.com/epiforecasts/mme-review

References

1. Green LE, Medley GF. Mathematical modelling of the foot and mouth disease epidemic of 2001: strengths and weaknesses. Res Vet Sci. 2002 Dec;73(3):201–5.

2. Hollingsworth TD, Medley GF. Learning from multi-model comparisons: Collaboration leads to insights, but limitations remain. Epidemics. 2017 Mar 1;18:1–3.

3. Reich NG, Lessler J, Funk S, Viboud C, Vespignani A, Tibshirani RJ, et al. Collaborative Hubs: Making the Most of Predictive Epidemic Modeling. Am J Public Health. 2022 Jun;112(6):839–42.

4. Teerawattananon Y, Kc S, Chi YL, Dabak S, Kazibwe J, Clapham H, et al. Recalibrating the notion of modelling for policymaking during pandemics. Epidemics. 2022 Mar;38:100552.

5. Shea K, Runge MC, Pannell D, Probert WJM, Li SL, Tildesley M, et al. Harnessing multiple models for outbreak management. Science. 2020 May 8;368(6491):577–9.

6. Medley GF. A consensus of evidence: The role of SPI-M-O in the UK COVID-19 response. Adv Biol Regul. 2022 Dec;86:100918.

7. Bates JM, Granger CWJ. The Combination of Forecasts. OR. 1969;20(4):451–68.

8. Makridakis S, Hyndman RJ, Petropoulos F. Forecasting in social settings: The state of the art. Int J Forecast. 2020 Jan 1;36(1):15–28.

9. Chen L. A review of the applications of ensemble forecasting in fields other than meteorology. Weather. 2024;79(9):285–90.

10. Claeskens G, Magnus JR, Vasnev AL, Wang W. The forecast combination puzzle: A simple theoretical explanation. Int J Forecast. 2016 Jul 1;32(3):754–62.

11. Del Valle SY, McMahon BH, Asher J, Hatchett R, Lega JC, Brown HE, et al. Summary results of the 2014-2015 DARPA Chikungunya challenge. BMC Infect Dis. 2018 May 30;18(1):245.

12. Ray E. Challenges in training ensembles to forecast COVID-19 cases and deaths in the United States [Internet]. International Institute of Forecasters. 2021 [cited 2021 Aug 5]. Available from: https://forecasters.org/blog/2021/04/09/challenges-in-training-ensembles-to-forecast-covid-19-cases-and-deaths-in-the-united-states/

13. Fox SJ, Kim M, Meyers LA, Reich NG, Ray EL. Optimizing Disease Outbreak Forecast Ensembles. Emerg Infect Dis. 2024 Sep;30(9):1967–9.

14. Kucharski AJ, Funk S, Eggo RM. The COVID-19 response illustrates that traditional academic reward structures and metrics do not reflect crucial contributions to modern science. PLOS Biol. 2020 Oct 16;18(10):e3000913.

15. Pollett S, Johansson M, Biggerstaff M, Morton LC, Bazaco SL, Brett Major DM, et al.

Identification and evaluation of epidemic prediction and forecasting reporting guidelines: A systematic review and a call for action. Epidemics. 2020 Dec 1;33:100400.

16. Scells H, Zuccon G. searchrefiner: A Query Visualisation and Understanding Tool for Systematic Reviews. In: Proceedings of the 27th International CIKM Conference on Information and Knowledge Management. 2018. p. 1939–42.

17. van de Schoot R, de Bruin J, Schram R, Zahedi P, de Boer J, Weijdema F, et al. An open source machine learning framework for efficient and transparent systematic reviews. Nat Mach Intell. 2021 Feb;3(2):125–33.