8. Susilawathi NM, Suryapraba AA, Soejitno A, Asih MW, Swastika K, Wandra T, et al. Neurocysticercosis cases identified at Sanglah Hospital, Bali, Indonesia from 2014 to 2018. Acta Trop. 2020;201:105208. https://doi.org/10.1016/j.actatropica.2019.105208
9. Smyth J, Adams V, Napier S. Case report: getting it taped. Br Dent J. 2015;219:146. https://doi.org/10.1038/sj.bdj.2015.639
10. Australian Centre for International Agricultural Research. Evaluating the opportunities for smallholder livestock keepers in Timor-Leste—final report [cited 2024 May 1]. https://www.aciar.gov.au/publication/LS-2017-035-final-report

Address for correspondence: Sung Hye Kim, Department of Environmental Biology and Medical Parasitology, Institute for Health and Society, Hanyang University College of Medicine, 222 Wangshipni-ro, Seongdong-gu, Seoul 04763, South Korea; email: sunghyekim@hanyang.ac.kr

Optimizing Disease Outbreak Forecast Ensembles

Spencer J. Fox, Minsu Kim, Lauren Ancel Meyers, Nicholas G. Reich, Evan L. Ray

Author affiliations: University of Georgia, Athens, Georgia, USA (S.J. Fox); University of Massachusetts Amherst, Amherst, Massachusetts, USA (M. Kim, N.G. Reich, E.L. Ray); University of Texas at Austin, Austin, Texas, USA (L.A. Meyers); Dell Medical School, Austin (L.A. Meyers); Santa Fe Institute, Santa Fe, New Mexico, USA (L.A. Meyers)

DOI: https://doi.org/10.3201/eid3009.240026

On the basis of historical influenza and COVID-19 forecasts, we found that more than 3 forecast models are needed to ensure robust ensemble accuracy. Additional models can improve ensemble performance, but with diminishing accuracy returns. This understanding will assist with the design of current and future collaborative infectious disease forecasting efforts.

Real-time collaborative forecast efforts have become the standard to generate and evaluate forecasts for infectious disease outbreaks (*1*,*2*). Individual forecasts are aggregated into an ensemble prediction that has historically outperformed individual models and is the primary external communication used (*3*–*5*). Because of the focus on the singular ensemble model and the costs associated with producing individual forecasts, public health officials starting or maintaining a forecast hub face 2 key challenges: identifying target participation rates and optimizing ensemble performance of participating models. To guide this decision-making, we analyzed data from recent US-based collaborative outbreak forecast hubs to identify how the size and composition of an ensemble influences performance.

We analyzed hub forecasts for influenza-like illness (ILI) from 2010–2017 (*5*); for COVID-19 reported cases, hospital admissions, and deaths from 2020–2023 (*6*); and for influenza hospital admissions from 2021–2023 (*7*). For each hub, we identified time periods with maximal model participation that had at least 2 increasing and 2 decreasing epidemiologic phases and obtained forecasts for individual models that produced ≥90% of all possible forecasts (Appendix Table 1, Figure 1, https://wwwnc.cdc.gov/EID/article/30/9/24-0026-App1.pdf). For each ensemble size, $n_D \in \{1, \dots, N_D\}$, where N_D is the disease-specific total number of models matching our inclusion criteria; we created unweighted ensemble forecasts for every combination of individual models of size n_D. We followed the hub forecast methodologies and made probabilistic forecasts for ILI by using a linear pool methodology (*5*), and we made quantile forecasts for all others by taking the median across all individual forecasts (Figure 1) (*8*). For each hub, we compared the ensemble performance against 2 hub-produced models. The first is a baseline model that produces naive forecasts and serves as a skill reference point; and the second is the published ensemble produced in real-time that is an unweighted ensemble of all submitted forecasts and is the current standard for performance (*3*,*5*). We summarized probabilistic ensemble forecast skill by using the log score for ILI forecasts and the weighted interval score for all others (*9*,*10*). We took the reciprocal of the log score so that lower values would indicate better performance similar to the weighted interval score (Appendix).

Looking across all ensemble sizes and combinations, we found that including more models improved average forecast performance and that all ensembles composed of >3 models outperformed the baseline model (Figure 2). Further increases to the ensemble size slightly improved the average forecast performance, but substantially decreased the variability of performance across ensembles. When we increased the ensemble size of influenza hospital admission forecasts from 4 to 7, the average

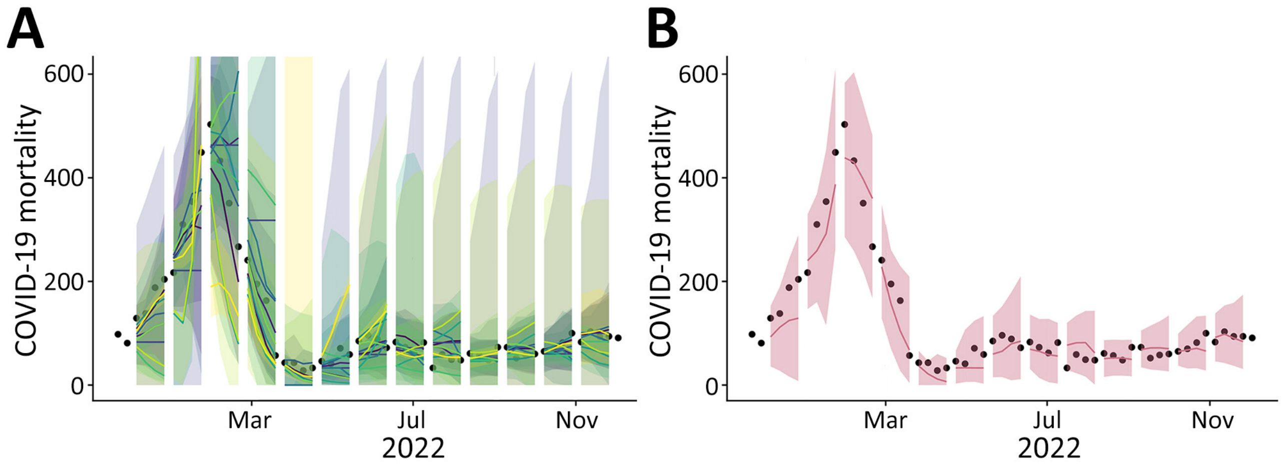

Figure 1. Comparison between individual and ensemble forecasts for COVID-19 mortality for Massachusetts, USA, from 1–4 weeks ahead, November 15, 2021–December 3, 2022, in study of optimizing disease outbreak forecasting ensembles. A) Individual forecasts of 10 models meeting inclusion criteria compared with weekly COVID-19 mortality estimates. B) An ensemble forecast constructed by taking the median across 10 individual forecasts compared with weekly COVID-19 mortality estimates. Black dots, weekly COVID-19 mortality estimates; colored lines, medians; ribbons, 95% prediction intervals.

performance improved by 2%, but the interquartile range decreased by 56.5%. Increasing the ensemble size therefore reduces the variability in expected performance of an ensemble.

To assist with decision-making regarding optimal ensemble assembly, we tested 2 approaches for model selection on the basis of past performance. We either ranked models by their individual performance and chose the top n_D models (individual rank) or we compared the performance of all ensemble combinations of size n_D and chose the models from the top performing ensemble (ensemble rank). Across all hubs, the individual rank methodology outperformed randomly assembled ensembles of

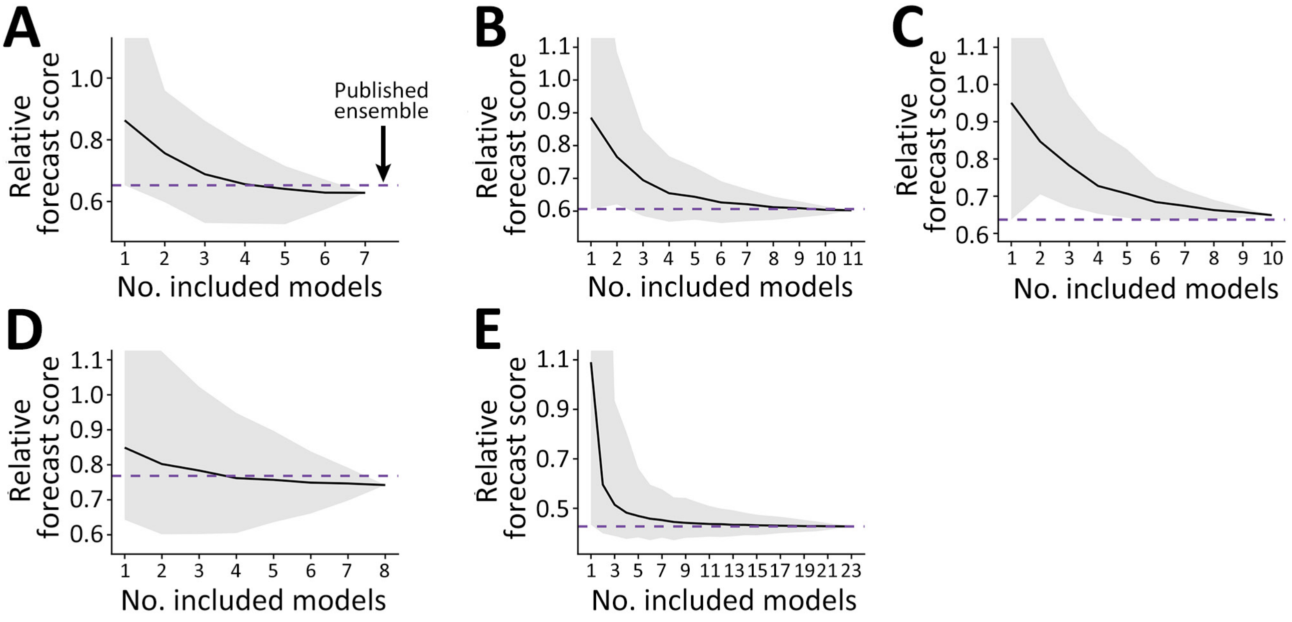

Figure 2. Summarized ensemble forecast scores from collaborative forecast efforts in study of optimizing disease outbreak forecasting ensembles. Scores correspond to the average forecast performance during testing periods across all dates, locations, and forecast horizons (Appendix Table 1, https://wwwnc.cdc.gov/EID/article/30/9/24-0026-App1.pdf). All scores are standardized by the baseline forecast model for that metric (Y = 1). Scores <1 indicate better accuracy than baseline. A) COVID-19 cases with 15 included models. B) COVID-19 admissions with 17 included models. C) COVID-19 deaths with 19 included models. D) Influenza admissions with 21 included models. D) Influenza-like illness with 23 included models. Solid black lines indicate mean scores; gray shading indicates minimum–maximum range. Horizontal purple dashed line indicates unweighted published ensemble used as standard.

the same size 63% (range 33.1%–87.2%) of the time, and the ensemble rank methodology outperformed randomly assembled ensembles of the same size 87.9% (range 70.9%–99.7%) of the time (Appendix Table 2, Figure 2). Performance of those ensembles is similar during both the training and testing periods, suggesting that ensemble performance is consistent through time (Appendix Figures 2, 3). Overall, ensemble rank outperforms individual rank for ensemble construction for 89.8% (range 66.7%–100%) of all sizes, and it provides a 6.1% (range 1.3%–11.9%) skill improvement (Appendix Table 2). The size 4 ensemble rank performed similarly to the published hub ensemble, although performance often declined with additional models (Appendix Figures 2, 3). Relative forecast performance across ensemble strategies was consistent when stratified by the ensemble size, forecast location, forecast date and phase of the epidemic, forecast target, and the skill metric (Appendix Figures 4–18).

Our results provide guidance for future collaborative forecast efforts. Hub organizers should target a minimum of 4 validated forecast models to ensure robust performance compared with baseline models. Adding more models reduces the variability in expected ensemble performance but might come with diminishing returns in average forecast skill. Organizers should use past ensemble performance rather than individual performance when selecting models to include in forecast ensembles; it is likely that further gains and different relationships between ensemble size and performance will come from weighted ensemble approaches (*8*). As public health officials and researchers look to expand collaborative forecast efforts, and as funding agencies allocate budgets across methodological and applied forecast efforts, our results can be used to identify target participation rates, assemble appropriate forecast models, and further improve ensemble forecast performance.

Acknowledgments

We thank the Council for State and Territorial Epidemiologists, Centers for Disease Control and Prevention, the Models of Infectious Disease Agent Study forecasting working groups, and the Scenario Modeling Hub. We also thank the Texas Advanced Computing Center at The University of Texas at Austin for providing high performance computing resources.

Funding for S.J.F. and L.A.M. was provided by the Council for State and Territorial Epidemiologists (grant no. NU38OT000297) and the Centers for Disease Control and Prevention (grant no. 75D30122C14776). Funding for M.K., E.L.R., and N.G.R. was provided by the National Institutes of General Medical Sciences (grant no. R35GM119582) and the Centers for Disease Control and Prevention (grant no. 1U01IP001122).

About the Author

Dr. Fox is an assistant professor at the University of Georgia in the department of epidemiology and biostatistics and the Institute of Bioinformatics. His research interests include statistical modeling of emerging infectious diseases and outbreak forecasting.

References

1. 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;112:839–42. https://doi.org/10.2105/AJPH.2022.306831
2. Biggerstaff M, Alper D, Dredze M, Fox S, Fung ICH, Hickmann KS, et al.; Influenza Forecasting Contest Working Group. Results from the Centers for Disease Control and Prevention's predict the 2013–2014 influenza season challenge. BMC Infect Dis. 2016;16:357. https://doi.org/10.1186/s12879-016-1669-x
3. Cramer EY, Ray EL, Lopez VK, Bracher J, Brennen A, Castro Rivadeneira AJ, et al. Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States. Proc Natl Acad Sci U S A. 2022;119:e2113561119. https://doi.org/10.1073/pnas.2113561119
4. Lutz CS, Huynh MP, Schroeder M, Anyatonwu S, Dahlgren FS, Danyluk G, et al. Applying infectious disease forecasting to public health: a path forward using influenza forecasting examples. BMC Public Health. 2019;19:1659. https://doi.org/10.1186/s12889-019-7966-8
5. Reich NG, Brooks LC, Fox SJ, Kandula S, McGowan CJ, Moore E, et al. A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States. Proc Natl Acad Sci U S A. 2019;116:3146–54. https://doi.org/10.1073/pnas.1812594116
6. Cramer EY, Huang Y, Wang Y, Ray EL, Cornell M, Bracher J, et al.; US COVID-19 Forecast Hub Consortium. The United States COVID-19 forecast hub dataset. Sci Data. 2022;9:462. https://doi.org/10.1038/s41597-022-01517-w
7. FluSight forecast -data 2022–2023 [cited 2023 Jul 12]. https://github.com/cdcepi/Flusight-forecast-data
8. Ray EL, Brooks LC, Bien J, Biggerstaff M, Bosse NI, Bracher J, et al. Comparing trained and untrained probabilistic ensemble forecasts of COVID-19 cases and deaths in the United States. Int J Forecast. 2023;39:1366–83. https://doi.org/10.1016/j.ijforecast.2022.06.005
9. Bracher J, Ray EL, Gneiting T, Reich NG. Evaluating epidemic forecasts in an interval format. PLOS Comput Biol. 2021;17:e1008618 https://doi.org/10.1371/journal.pcbi.1008618
10. Gneiting T, Raftery AE. Strictly proper scoring rules, prediction, and estimation. J Am Stat Assoc. 2007;102:359–78. https://doi.org/10.1198/016214506000001437

Address for correspondence: Spencer Fox, University of Georgia, 120 B.S. Miller Hall, Health Sciences Campus, 101 Buck Rd, Athens, GA 30602, USA; email: sjfox@uga.edu