Predicting seasonal influenza hospitalization using an ensemble super learner: a simulation study

Jason R. Gantenberg [1,2*], Kevin W. McConeghy [2,3], Laura B. Balzer [4], Chanelle J. Howe, [5], Andrew R. Zullo, [1,2]

1 Department of Epidemiology, Brown University School of Public Health, 121 S. Main St., Providence, RI, 02912
2 Department of Health Services, Policy and Practice, 121 S. Main St., Providence, RI, 02912
3 Providence VA Medical Center, 830 Chalkstone Ave., Providence, RI, 02908
4 Department of Biostatistics and Epidemiology, School of Public Health and Health Sciences, University of Massachusetts Amherst, 427 Arnold House, 715 N. Pleasant St., Amherst, MA 01003
5 Department of Epidemiology, Center for Epidemiology and Environmental Health, Brown University School of Public Health, 121 S. Main St., Providence, RI, 02912

* Corresponding author: jrgant@brown.edu

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Curabitur eget porta erat. Morbi consectetur est vel gravida pretium. Suspendisse ut dui eu ante cursus gravida non sed sem. Nullam sapien tellus, commodo id velit id, eleifend volutpat quam. Phasellus mauris velit, dapibus finibus elementum vel, pulvinar non tellus. Nunc pellentesque pretium diam, quis maximus dolor faucibus id. Nunc convallis sodales ante, ut ullamcorper est egestas vitae. Nam sit amet enim ultrices, ultrices elit pulvinar, volutpat risus.

Author summary

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Curabitur eget porta erat. Morbi consectetur est vel gravida pretium. Suspendisse ut dui eu ante cursus gravida non sed sem. Nullam sapien tellus, commodo id velit id, eleifend volutpat quam. Phasellus mauris velit, dapibus finibus elementum vel, pulvinar non tellus. Nunc pellentesque pretium diam, quis maximus dolor faucibus id. Nunc convallis sodales ante, ut ullamcorper est egestas vitae. Nam sit amet enim ultrices, ultrices elit pulvinar, volutpat risus.

Meta

Target journal: *PLoS Computational Biology*
Section: Epidemiology and Clinical/Translational Studies
Potential editors:

- Benjamin Althouse
- Miles Davenport

- Matthew Ferrari
- Roger Kouyos
- James Lloyd-Smith

Potential reviewers:

- Ryan J. Tibshirani (co-author on paper we use for curve simulation)
- Logan C. Brooks (co-author on paper we use for curve simulation)
- Roni Rosenfeld (co-author on paper we use for curve simulation)
- Sherri Rose
- David A. Osthus
- Samrachana Adhikari

Introduction

Each year, seasonal influenza causes approximately XXXX hospitalizations and XXXX deaths per year in the United States alone [cite]. Being able to predict how influenza-related hospitalizataions will change over time during any given influenza season can assist policymakers, public health officials, and physicians allocate resources appropriately and prepare more efficiently for changes in hospitalization rates.

While influenza forecasting is a still-maturing science [1], researchers have made considerable progress over the past decade in improving the quality of and capacity for forecasting influenza-like illness (ILI) [cite], thanks in part to the FluSight forecasting competitions sponsored by the Centers for Disease Control and Prevention (CDC) since the 2013–14 flu season [cite]. Many different types of models have been used to generate forecasts, including statistical time series models [1,2], Bayesian regression [cite], and agent-based models [cite], among others. However, ensemble methods have emerged as perhaps the most promising approach to improving the accuracy and stability of epidemic predictions [3,4].

Ensembles combine predictions generated by a set of component models [3,5–7]. In some cases, ensembles aggreggate component model predictions by weighting better predictions more highly in the final ensemble prediction [3,4], though other weighting criteria can be applied [4]. The rationale for using ensemble predictions rests in their ability to borrow the strengths and discard the weaknesses of various component models. This feature tends to lead not only to more accurate predictions but to more stable ones that can be applied across a range of scenarios [4]. The CDC's primary in-season ILI forecasts are now based on an ensemble forecast generated by aggregating predictions from a growing library of individual forecasts submitted by research teams around the U.S. [].

To date, most work has focused on ILI [1,2,8–10], with considerably less effort having been exerted so far on predicting influenza-related hospitalization rates [11]. Because the dynamics of flu-related hospitalizations might evolve differently over the course of an influenza season—at the very least, lagging influenza incidence by a week or two [citation needed]—and because hospitalization rates are an independent signal of the severity of disease caused by circulating flu strains, optimizing ensembles to predict hospitalization rates can provide complementary information to ILI forecasts.

One ensemble machine learning method in particular, dubbed "super learner" [12–14], exhibits a number of desirable properties that suggest it may be a powerful tool for predicting flu hospitalizations. First, its developers have demonstrated that, asymptotically, the super learner is an oracle estimator, performing as well as the best-fitting component model and converging almost as quickly [13] [also will want to read and cite the 2003 paper of van der Laan's]. Second, this oracle property generally translates to finite samples [cite correct Polley and van der Laan papers]. Finally, several packages have been developed to implement the super learner algorithm [15,16], providing researchers easy access to a relatively large library of component models and a means to calculate cross-validated prediction risks quite easily [16].

In this study, we sought to train an ensemble learner on a distribution of simulated influenza hospitalization curves to generate predictions for three seasonal target parameters based on the CDC forecasting competitions [17]: peak hospitalization rate, peak week of the season, and cumulative hospitalization rate. We sought to compare the performance of the ensemble learner to the best-performing component model and a naive historical average prediction across the 30 weeks of a flu season for each of these three prediction targets.

Methods

Results

Discussion

Software and code

All code is provided at … [set up persistent DOI at Zenodo or Open Science Framework and link to Github repo for FluHospPrediction package]

Declarations

Acknowledgement

We thank Ashley Naimi and Nicholas Reich for their comments on earlier versions of the analysis plan.

Funding statement

This work was funded by an unrestricted grant from Sanofi (PI: Andrew Zullo). The funders did not assist in the statistical analysis nor did they have a say in the final decision to submit the manuscript for publication.

Competing interests

[solicit competing interests from co-authors]

References

Supporting information

1. 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: 3146–3154. doi:10.1073/pnas.1812594116

2. Biggerstaff M, Kniss K, Jernigan DB, Brammer L, Bresee J, Garg S, et al. Systematic assessment of multiple routine and near Real-Time indicators to classify the

severity of influenza seasons and pandemics in the united states, 2003-2004 through 2015-2016. Am J Epidemiol. 2018;187: 1040–1050. doi:10.1093/aje/kwx334

3. Reich NG, McGowan CJ, Yamana TK, Tushar A, Ray EL, Osthus D, et al. Accuracy of real-time multi-model ensemble forecasts for seasonal influenza in the U.S. PLoS Comput Biol. 2019;15: e1007486. doi:10.1371/journal.pcbi.1007486

4. Ray EL, Reich NG. Prediction of infectious disease epidemics via weighted density ensembles. PLoS Comput Biol. 2018;14: e1005910. doi:10.1371/journal.pcbi.1005910

5. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: Data mining, inference, and prediction. Springer, New York, NY; 2009. doi:10.1007/978-0-387-84858-7

6. Wolpert DH. Stacked generalization. Neural Netw. 1992;5: 241–259. doi:10.1016/S0893-6080(05)80023-1

7. Breiman L. Stacked regressions. Mach Learn. 1996;24: 49–64. doi:10.1007/BF00117832

8. McGowan CJ, Biggerstaff M, Johansson M, Apfeldorf KM, Ben-Nun M, Brooks L, et al. Collaborative efforts to forecast seasonal influenza in the united states, 2015-2016. Sci Rep. nature.com; 2019;9: 683. doi:10.1038/s41598-018-36361-9

9. Kandula S, Yamana T, Pei S, Yang W, Morita H, Shaman J. Evaluation of mechanistic and statistical methods in forecasting influenza-like illness. J R Soc Interface. 2018;15. doi:10.1098/rsif.2018.0174

10. Brooks LC, Farrow DC, Hyun S, Tibshirani RJ, Rosenfeld R. Flexible modeling of epidemics with an empirical bayes framework. PLoS Comput Biol. 2015;11: e1004382. doi:10.1371/journal.pcbi.1004382

11. Kandula S, Pei S, Shaman J. Improved forecasts of influenza-associated hospitalization rates with google search trends. J R Soc Interface. 2019;16: 20190080. doi:10.1098/rsif.2019.0080

12. Laan MJ van der, Polley EC, Hubbard AE. Super learner. Stat Appl Genet Mol Biol. De Gruyter; 2007;6: Article25. doi:10.2202/1544-6115.1309

13. Polley EC, Laan MJ van der. Super learner in prediction. University of California, Berkeley; 2010.

14. Polley EC, Rose S, Laan MJ van der. Super learning. In: Laan MJ van der, Rose S, editors. Targeted learning: Causal inference for observational and experimental data. New York, NY: Springer New York; 2011. pp. 43–66. doi:10.1007/978-1-4419-9782-1_3

15. Polley E, LeDell E, Kennedy C, van der Laan M. SuperLearner: Super learner prediction [Internet]. 2019. Available: https://CRAN.R-project.org/package=SuperLearner

16. Coyle JR, Hejazi NS, Malenica I, Sofrygin O. Sl3: Pipelines for machine learning and super learning. 2020. doi:10.5281/zenodo.1342293

17. Centers for Disease Control and Prevention. Epidemic prediction initiative. https://predict.cdc.gov/post/59973fe26f7559750d84a843;