JAMA Network Open

Invited Commentary | Health Informatics

Machine Learning for Prediction in Electronic Health Data

Sherri Rose, PhD

+ Related article

Author affiliations and article information are listed at the end of this article.

Machine learning for prediction in electronic health data has been deployed for many clinical questions during the last decade. Machine learning methods may excel at finding new features or nonlinear relationships in the data, as well as handling settings with more predictor variables than observations. However, the usefulness of both these data and machine learning has varied. Electronic health data often have quality issues (eg, missingness, misclassification, measurement error), and machine learning may perform similarly to standard techniques for some research questions. Ensembles (running multiple algorithms and either selecting the single best algorithm or creating a weighted average) can help mitigate the latter concern. Using several machine learning tools, Wong et al[1] predicted delirium risk for newly hospitalized patients with high-dimensional electronic health record data at a large academic health institution. They compared these approaches with a questionnaire-based scoring system and found improved performance for machine learning with respect to several metrics calculated in a single holdout sample. Their article is a step toward updating delirium risk prediction. It also provides an opportunity to discuss 2 key issues in the current state of machine learning for prediction in electronic health data: evaluation and generalizability.

The machine learning researchers who develop novel algorithms for prediction and the clinical teams interested in implementing them are frequently and unfortunately 2 nonintersecting groups. Thus, these algorithms may originally be built and evaluated in the machine learning literature based on metrics that are less clinically useful than other choices. Computer scientists and statisticians may optimize to achieve the best area under the receiver operating characteristic curve (AUC), but what a clinical team might need is high sensitivity or positive predictive value. Worse yet, algorithms can have misleading performance when evaluated only along 1 or 2 dimensions. For example, high AUC, accuracy, and positive predictive values can be accompanied by near-zero levels of sensitivity and specificity. It can also be essential to calculate metrics like the percentage of true cases in the top risk percentile.[2] This is especially important when a goal of the tool is to target high-risk patients for interventions. Wong and colleagues[1] helpfully compute AUC, positive predictive values, sensitivity, and specificity. However, one accepted standard for evaluation that should be adopted in future clinical machine learning applications is *K*-fold cross-validation (eg, with $K = 10$) rather than the single holdout validation sample in their article. *K*-fold cross-validation involves *K* successive mutually exclusive validation sets where the algorithm fitting is iteratively performed on the nonvalidation data (ie, training set). At the end, each observation in the full data has a predicted value that was obtained from when it was part of a validation set. Typically, metrics calculated based on these *K*-fold cross-validated predicted values will more effectively assess overfitting and have lower variance.[3]

Of course, assessing the generalizability of a prediction algorithm goes well beyond using cross-validated metrics to evaluate overfitting. Wong et al[1] carefully discussed a number of limitations in their work, including the lack of external validation in other health systems. As in similar studies, the step from good performance in the study to a generalizable algorithm is vast and sometimes may not be feasible. Patients receiving treatment in varied care settings or geographic regions simply may require tailored tools. Recognizing the need for unique tools in different populations is not inherently negative, but one of many considerations not magically solved by using machine learning. Even those algorithms that prove to be generalizable may quickly become outdated as treatment patterns or physician incentives to code health conditions change.[4] Increased social tolerance for certain

conditions, such as mental health and substance use disorders, may also necessitate updating tools. How often algorithms should be reevaluated will depend on many of these factors.

When implementing machine learning for prediction in electronic health data, it is critical to remember that these data are not collected to answer specific research questions, which is a central difficulty in relying on them for these purposes. Machine learning might or might not provide benefits and the data might not be robust enough to be useful to clinical teams. Prediction algorithms that combine the best clinical expertise with rigorous machine learning tools are the most promising for continued work.[5,6] Wong et al[1] brought this to their study by using an expert clinical panel to select the 796 predictor variables available to their machine learning algorithms, while still allowing for automated variable selection. As Wong and colleagues[1] also demonstrated, it is imperative to include a comparison to standard practice. Their top-performing algorithm improved nontrivially over the questionnaire-based scoring system, although some metrics were still low (59.7% sensitivity and 23.1% positive predictive value). Whether these levels are high enough to support use deserves further discussion. In general, algorithms can be designed with a priori thresholds for both improvement over current practice and the minimum level needed along each metric to warrant investing in establishing a new prediction system.

There is reason to be optimistic about the ability of machine learning to transform prediction in an array of medical fields. Machine learning also has demonstrated promise in clinical domains when the goal is to discover clusters in the data, such as imaging analysis for therapeutic selection.[7] Here, the new features can be validated with expert evaluation from radiologists or neurologists, which differs from the prediction setting where observed labels exist in the data. Causal inference methods that incorporate machine learning[8] are a burgeoning area, including techniques for treatment effect heterogeneity.[9] Understanding heterogeneous treatment effects will likely be one path in the journey toward precision medicine. That said, we are in a discovery phase, and the pervasiveness of electronic health big data across many clinical areas does not ultimately mean machine learning will be equally valuable in each.

ARTICLE INFORMATION

Published: August 3, 2018. doi:10.1001/jamanetworkopen.2018.1404

Corresponding Author: Sherri Rose, PhD, Department of Health Care Policy, Harvard Medical School, 180 Longwood Ave, Boston, MA 02115 (rose@hcp.med.harvard.edu).

Author Affiliation: Department of Health Care Policy, Harvard Medical School, Boston, Massachusetts.

Conflict of Interest Disclosures: Dr Rose reported grants from the National Institutes of Health (NIH) during the conduct of the study.

Funding/Support: Research reported in this publication was supported by the NIH through an NIH Director's New Innovator Award DP2-MD012722.

Role of the Funder/Sponsor: The NIH had no role in the preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Disclaimer: The content is solely the responsibility of the author and does not necessarily represent the official views of the NIH.

REFERENCES

1. Wong A, Young AT, Liang AS, Gonzales R, Douglas VC, Hadley D. Development and validation of an electronic health record–based machine learning model to estimate delirium risk in newly hospitalized patients without known cognitive impairment. *JAMA Netw Open*. 2018;1(4):e181018. doi:10.1001/jamanetworkopen.2018.1018

2. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*. 2010;21(1):128-138. doi:10.1097/EDE.0b013e3181c30fb2

3. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proc Int Joint Conf Artif Intell*. 1995;14(2):1137-1145.

4. Bergquist S, Brooks G, Keating N, Landrum MB, Rose S. Classifying lung cancer severity with ensemble machine learning in health care claims data. http://proceedings.mlr.press/v68/bergquist17a.html. *Proc Mach Learn Res*. 2017;68:25-38.

5. Verghese A, Shah NH, Harrington RA. What this computer needs is a physician: humanism and artificial intelligence. *JAMA*. 2018;319(1):19-20. doi:10.1001/jama.2017.19198

6. Beam AL, Kohane IS. Big data and machine learning in health care. *JAMA*. 2018;319(13):1317-1318. doi:10.1001/jama.2017.18391

7. Sweeney EM, Shinohara RT, Dewey BE, et al. Relating multi-sequence longitudinal intensity profiles and clinical covariates in incident multiple sclerosis lesions. *Neuroimage Clin*. 2015;10:1-17. doi:10.1016/j.nicl.2015.10.013

8. van der Laan M, Rose S. *Targeted Learning: Causal Inference for Observational and Experimental Data*. New York, NY: Springer; 2011. doi:10.1007/978-1-4419-9782-1

9. Athey S, Imbens G. Recursive partitioning for heterogeneous causal effects. *Proc Natl Acad Sci U S A*. 2016;113(27):7353-7360. doi:10.1073/pnas.1510489113