HHS Public Access

Author manuscript

Circ Cardiovasc Qual Outcomes. Author manuscript; available in PMC 2022 May 23.

Published in final edited form as:

Circ Cardiovasc Qual Outcomes. 2020 October ; 13(10): e007491. doi:10.1161/CIRCOUTCOMES.120.007491.

Machine Learning in Clinical Journals: Moving from Inscrutable to Informative

Karandeep Singh, MD, MMSc[1,2,3,4], **Andrew L. Beam, PhD**[5], **Brahmajee K. Nallamothu, MD, MPH**[2,4]

[1]Department of Learning Health Sciences, University of Michigan Medical School, Ann Arbor, MI

[2]Department of Internal Medicine, University of Michigan Medical School, Ann Arbor, MI

[3]School of Information, University of Michigan, Ann Arbor, MI

[4]Michigan Integrated Center for Healthcare Analytics and Medical Prediction, Ann Arbor, MI

[5]Department of Epidemiology at the Harvard T.H. Chan School of Public Health, Boston, MA

Keywords

machine learning; reporting guidelines

Although machine learning (ML) algorithms have grown more prevalent in clinical journals, inconsistent reporting of methods has led to skepticism about ML results and has blunted their adoption into clinical practice. A common problem authors face when reporting on ML methods is the lack of a single reporting guideline that applies to the panoply of ML problem types and approaches. Responding to this concern, Stevens et. al. in this issue of *Circulation: Cardiovascular Quality & Outcomes* propose a set of reporting recommendations for ML papers.[1] This is meant to augment existing (but more general) reporting guidelines on predictive models, such as the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) statement[2] – recognizing that international efforts are underway to address these concerns more fully, including TRIPOD-ML and CONSORT-AI.[3,4] However, a more fundamental question remains: will such guidelines address all of the different types of ML methods increasingly used in the published literature (such as unsupervised learning) based on their proposed scope?

This is not a small problem in a dynamic field full of rapid change. While Stevens et. al. focus on the spectrum of ML problem types between unsupervised and supervised learning, for instance, more recent literature has also seen the application of reinforcement learning methods to clinical problems.[5,6] Additionally, the phrase “machine learning” can be used to refer to a number of different algorithms, which include decision trees, random

Corresponding author: Karandeep Singh, MD, MMSc, Department of Learning Health Sciences, University of Michigan Medical School, Ann Arbor, MI, Address: 1161H NIB, 300 N. Ingalls St., Ann Arbor, MI 48109, Phone: 734-936-1649, kdpsingh@umich.edu.

Disclosures: Disclosures provided by Brahmajee K. Nallamothu in compliance with American Heart Association’s annual Journal Editor Disclosure Questionnaire are available at https://www.ahajournals.org/pb-assets/COI_09-2019.pdf

forests, gradient-boosted machines, support vector machines, and neural networks. Perhaps somewhat confusingly, the phrase is also commonly applied to describe methods like linear regression and Bayesian models that readers of clinical journals are more familiar with. Thus, it is not uncommon to read a paper purport to use "machine learning" in the title, only to find the actual algorithm is a penalized regression model, which a statistical reviewer may consider to be a "traditional statistics model."

As with many types of clinical research, we believe standardized reporting of ML methods is a needed advancement and will be helpful for editors and reviewers evaluating the quality of manuscripts. Ultimately, this will be of great benefit to readers of the published article. In this editorial, however, we describe ongoing difficulties in determining what is and what is not an ML manuscript to which Stevens et. al.'s reporting recommendations would apply. We then discuss how this determination impacts how a paper is perceived and judged by journals and readers. On the one hand, strong model performance described in papers may be the result of overfitting, and this could be prevented with better reporting practices. On the other hand, valid ML findings may be overlooked by even experienced statistical editors due to lack of familiarity with specific methods. Although the Stevens et. al. recommendations provide an important first step towards improving ML papers, we propose that clinical journals also may benefit from creating ML editorship roles (as *Circulation: Cardiovascular Quality and Outcomes* is doing) to work alongside statistical editors.

What's in a Name?

The wide range of ML problem types and approaches highlights a deeper and much more difficult-to-solve taxonomy challenge: what exactly is meant by the phrase "machine learning" in clinical journals. Furthermore, the machine learning community is heterogenous and comprises researchers with various backgrounds ranging from computer science, informatics, mathematics, operations engineering, and statistics. Given the interdisciplinary nature of machine learning, it can be difficult to understand how a proposed method may relate to traditional statistical models commonly found in the clinical literature. This complicates the review process as it is difficult to evaluate whether a proposed method constitutes a genuinely new advance. Stevens et. al. try to draw contrasts between ML methods and "traditional statistical methods," which they sometimes refer to as "hypothesis-driven approaches." This contrast is intended to help readers of clinical studies – who are often more familiar with traditional statistical methods – determine whether a given paper uses ML methods and thus is subject to their reporting recommendations.

However, reasonable scientists may disagree (and often do!) on which approaches fall into the umbrella terms of "ML" and statistics, and much of this disagreement stems from the complicated history surrounding the discovery and use of these methods.[7–9] For instance, lasso regression was popularized by a statistician but is used more commonly by ML practitioners (and sometimes referred to as "L1-regularization").[10] Similarly, the random forest algorithm was first proposed by a computer scientist but popularized and trademarked by two statisticians and yet is not considered a traditional statistics method. Even the use of hypothesis testing does not reliably determine that an algorithm belongs to traditional

statistics. Conditional inference trees use a hypothesis test to determine optimal splits but are considered ML due to their resemblance to random forests.[11]

For this reason, we caution against the use of the phrase “machine learning” in clinical journals in isolation in the title. Instead, we suggest naming the specific algorithm used in the title (e.g., “K-means clustering” or “penalized regression”) or omitting the method from the title in favor of a more general term (e.g., “clustering method” or “prediction model”). If there is ambiguity in the problem type after naming the algorithm (which is exceedingly rare), then we suggest labeling it as a supervised or unsupervised learning task. This precision will allow readers to be in a better position to judge papers on the appropriate application of the algorithm as opposed to focusing on whether the paper is, in fact, ML.

Stevens et. al. also suggest that authors provide a rationale for the use of ML methods in place of traditional statistics methods. As when any method is applied, we see value in describing to readers the reasoning for applying a specific tool. Yet even a statement of rationale could invoke seemingly partisan divides between ML and statistics reviewers as the best approach may depend on the signal-to-noise ratio of the problem and may not be knowable in advance. On one hand, an independent evaluation of hundreds of algorithms on over a hundred publicly available datasets found ML algorithms to perform substantially better than logistic regression.[12] On the other hand, two systematic reviews of the medical literature found no benefit of machine learning methods over logistic regression.[13,14] These differences could result from the lower signal-to-noise ratio in biomedical data versus other types of data, publication bias, or model misspecification.

Better Reporting May Uncover Poor Modeling Practices

There are many situations, particularly with high-dimensional temporal (e.g., clinical time series) or spatial data (e.g., imaging), where ML methods outperform traditional statistical methods. Unfortunately, many common errors in the application of supervised ML methods lead to overestimation of model performance. For example, imputing data or applying data-driven feature selection methods *before* splitting the training and test sets can lead to overestimates in model performance. This phenomenon occurs relatively commonly and has been referred to as the “Winner’s curse” or “testimation bias.”[15] Similarly, using cross-validation to select hyperparameters in the absence of nesting can lead to overestimates of internal validity. Finally, use of billing codes as predictors in electronic health record data may lead to favorably biased results due to delayed data entry because data provided to the model retrospectively may not be available if the model were run prospectively. Because clinical journals may be more likely to consider papers with strong model performance (e.g., high area-under-the-curve), peer reviewers may be disproportionately asked to review papers with unrecognized overfitting.

Uncovering concerns like overfitting requires authors to be clear in their methods. Stevens et. al. identify a number of important considerations relevant to ML papers that deserve to be highlighted during reporting, including data quality, feature selection, model optimization (or tuning), and code sharing. Improving the reporting of ML algorithms in these aspects will improve the quality of clinical ML papers by enabling editors, reviewers, and readers to

appropriately scrutinize and improve the quality of the work. However, these safeguards may not be sufficient for all concerns. For example, none of these would prevent the widespread use of models with substantial yet often unrecognized problems such as racial bias.[16,17]

While problems like racial bias are more apparent when the models are transparent (such linear regression models or decision trees), the use of more opaque methods can obscure them. Interpreting the learned relationships between predictors and the outcome can help determine when a model may be learning an unintended representation. Although this is commonly done using permutation-based methods or class activation maps,[18,19] even these methods may not provide an accurate picture of what the model has learned.[20,21] Thus, use of model interpretability methods may instill a false sense of confidence in readers and should be reported with caution.

Sharing model objects can also facilitate independent evaluations of ML models — these objects are files that contain a representation of the model itself. Sharing of model objects is common in ML literature but rarely done in clinical journals.[22–24] For example, neural network weights may be shared in “hdf5,” “pt,” or “mojo” file formats for models trained using the keras, PyTorch, and h2o packages, respectively. Another common way to save and share ML models is in native R (“rds”) or Python (“pickle”) binary files. Standard formats do exist for sharing ML models such as the Predictive Model Markup Language but limited integration with ML software packages have prevented widespread use.[25,26] Few of these technical details are known to clinical researchers and model objects are rarely shared.

A Need for Machine Learning Editors at Clinical Journals

Statistical editors play a key role in ensuring that scientific research is conducted and reported in a sound manner. Faced with a manuscript using ML approaches, statistical editors may not always be in the best position to judge its scientific rigor. This is perhaps most true for research utilizing neural networks with complex architectures but applies to other algorithms as well. For instance, while there is helpful statistical guidance on the determination of appropriate sample sizes for prediction models,[27] it is unclear how this guidance could be applied to models utilizing radiology images consisting of millions of pixels. If each pixel were considered a predictor variable, almost every such paper would likely be flagged by a statistical editor as having an inadequate sample size. Now consider that an ML researcher may not refit the entire neural network on the radiology image data, deciding instead to reuse a model from a different domain and refitting only a subset of the parameters while freezing the rest of the model (an approach known as transfer learning). This approach can enable neural networks to be effectively fit on much smaller datasets through transfer of knowledge from larger to smaller datasets. Neural networks with a large number of parameters can be surprisingly robust on small datasets but this depends on the choice of model architecture, activation functions, hyperparameters (such as learning rate), and checks to ensure model convergence.[28]

These kinds of nuances may be identified through peer review, but clinical journals that increasingly deal with ML manuscripts may benefit from a more consistent approach to evaluating them. Stevens et. al. provide one piece of this puzzle through their reporting

recommendations but this approach needs to be augmented. The scientific quality of manuscripts reporting on ML models would be greatly improved with the creation of ML editorships, as some journals have already begun to do.[29] Like statistical editors focused in other areas, ML editors would be helpful in identifying peer reviewers with specific expertise in the ML algorithm being applied in a given manuscript and in ensuring that the description of the methods and results is accurate and appropriate for a clinical readership.

We have taken this approach at *Circulation: Cardiovascular Quality and Outcomes* where we have identified a new editor for handling ML manuscripts – cardiologist Dr. Rashmee Shah from the University of Utah – and also a new member of the statistical editorial team – computer scientist Dr. Bobak Mortazavi from the Texas A&M University. Combined with the reporting tools recommended by Stevens et al., we believe these two additions to our team will create a more rigorous and useful process for internally evaluating ML methods in studies submitted to us. In the end, these changes will help our readers better understand and apply findings from these exciting studies to our patients moving forward.

References

1. Stevens LM, Mortazavi BJ, Deo RC, Curtis L, Kao DK. Recommendations for reporting machine learning analyses in clinical research. Circ Cardiovasc Qual Outcomes. 2020;13:e006556.
2. Collins GS, Reitsma JB, Altman DG, Moons KGM, TRIPOD Group. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. The TRIPOD Group. Circulation. 2015;131:211–219. [PubMed: 25561516]
3. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. Lancet. 2019;393:1577–1579. [PubMed: 31007185]
4. Liu X, Faes L, Calvert MJ, Denniston AK, CONSORT/SPIRIT-AI Extension Group. Extension of the CONSORT and SPIRIT statements. Lancet. 2019;394:1225.
5. Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. Nat Med. 2018;24:1716–1720. [PubMed: 30349085]
6. Gottesman O, Johansson F, Komorowski M, Faisal A, Sontag D, Doshi-Velez F, Celi LA. Guidelines for reinforcement learning in healthcare. Nat Med. 2019;25:16–18. [PubMed: 30617332]
7. Breiman L. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). Statistical Science. 2001;16:199–231.
8. Finlayson S. Comments on ML “versus” statistics. https://sgfin.github.io/2020/01/31/Comments-ML-Statistics/. January 31, 2020. Accessed June 2, 2020.
9. Beam AL, Kohane IS. Big Data and Machine Learning in Health Care. JAMA. 2018;319:1317–1318. [PubMed: 29532063]
10. Tibshirani R. Regression shrinkage and selection via the lasso: a retrospective. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2011;73:273–282.
11. Hothorn T, Hornik K, Zeileis A. Unbiased Recursive Partitioning: A Conditional Inference Framework. Journal of Computational and Graphical Statistics. 2006;15:651–674.
12. Fernández-Delgado M, Cernadas E, Barro S, Amorim D. Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? J Mach Learn Res. 2014;15:3133–3181.
13. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol. 2019;110:12–22. [PubMed: 30763612]
14. Mahmoudi E, Kamdar N, Kim N, Gonzales G, Singh K, Waljee AK. Use of electronic medical records in development and validation of risk prediction models of hospital readmission: systematic review. BMJ. 2020;369:m958. [PubMed: 32269037]
15. Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. Springer, Cham; 2019.

16. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366:447–453. [PubMed: 31649194]
17. Grother P, Ngan M, Hanaoka K. Face Recognition Vendor Test (FVRT): Part 3, Demographic Effects. 2019.
18. Molnar C. Interpretable Machine Learning. 2020. https://christophm.github.io/interpretable-ml-book/index.html. Accessed June 3, 2020.
19. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning Deep Features for Discriminative Localization. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 2921–2929, doi: 10.1109/CVPR.2016.319.
20. Strobl C, Boulesteix A-L, Kneib T, Augustin T, Zeileis A. Conditional variable importance for random forests. BMC Bioinformatics. 2008;9:307. [PubMed: 18620558]
21. Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B. Sanity Checks for Saliency Maps. ArXiv. Preprint posted online October 28, 2018. http://arxiv.org/abs/1810.03292.
22. Open Neural Network Exchange. ONNX. http://onnx.ai. Accessed June 3, 2020.
23. Auffenberg GB, Ghani KR, Ramani S, Usoro E, Denton B, Rogers C, Stockton B, Miller DC, Singh K, Michigan Urological Surgery Improvement Collaborative. askMUSIC: Leveraging a Clinical Registry to Develop a New Machine Learning Model to Inform Patients of Prostate Cancer Treatments Chosen by Similar Men. Eur Urol. 2018;75:901–907. [PubMed: 30318331]
24. 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:e181018.
25. Grossman R, Bailey S, Ramu A, Malhi B, Hallstrom P, Pulleyn I, Qin X. The management and mining of multiple predictive models using the predictive modeling markup language. Information and Software Technology. 1999;41:589–595.
26. PMML 4.4 - General Structure. http://dmg.org/pmml/v4-4/GeneralStructure.html. Accessed June 9, 2020.
27. Riley RD, Ensor J, Snell KIE, Harrell FE Jr, Martin GP, Reitsma JB, Moons KGM, Collins G, van Smeden M. Calculating the sample size required for developing a clinical prediction model. BMJ. 2020;368:m441. [PubMed: 32188600]
28. You can probably use deep learning even if your data isn't that big. http://beamlab.org/deeplearning/2017/06/04/deep_learning_works.html. Published June 4, 2017. Accessed June 4, 2020.
29. Waljee appointed associate editor for the journal Gut. Institute for Healthcare Policy & Innovation. https://ihpi.umich.edu/news/waljee-appointed-associate-editor-journal-gut. Published February 15, 2020. Accessed June 3, 2020.