EDITORIAL

Evaluating Machine Learning Articles

Finale Doshi-Velez, PhD; Roy H. Perlis, MD, MSc

Related article page 1806

In this issue of *JAMA*, Liu and colleagues[1] provide a users' guide to reading clinical machine learning articles. Beyond a synopsis of selected concepts in modern machine learning, the authors elaborate step-by-step guidance for physicians seeking to evaluate this evidence with a critical eye. In an era when readers are bombarded with artificial intelligence in everyday life, from credit card fraud warnings and smartphones that anticipate their needs to life-like videos of people who do not actually exist, the sanity check provided by this article is most welcome.

Reassuringly, many of the key elements in reading a machine learning article draw directly on concerns familiar to *JAMA* readers of users' guides, and they have changed little in the 3 decades since Nierenberg described an approach to diagnostic testing.[2] Common sense and standard statistical principles still apply when it comes to these more complex models.

For example, choices about the inputs and outputs of a model, such as what and how patient features are measured and what is to be predicted, are essential in determining the practical value of an algorithm. Are the inputs measured reliably, and do they draw on readily available technology (facts from electronic health records; routine laboratory studies) or emerging technology (new positron emission tomography tracers, single-cell transcriptomics) that may make implementation and dissemination more challenging? Are the outputs clinically actionable? Generations of medical students recall the adage, "don't order a test unless it will change management"; certainly this applies to artificial intelligence as well. Tools to detect retinopathy[3] or identify tuberculosis or malaria using smartphone images[4] may be particularly beneficial in low-resource settings.

Choices about cohort selection and data preparation (most notably, handling of missing data) will have important consequences for subsequent analyses; machine learning does not solve problems of bias introduced by missing data. Were models trained only with canonical or clear-cut examples? In clinical practice, data are noisy and not always complete; failure to consider these circumstances may yield models that perform beautifully on cleaned data sets for the purposes of publication but miserably in practice. Radiologists do not struggle to identify cancer in pristine chest radiologic images accompanied by detailed history but poorer-quality images with superimposed pneumonia and little clinical context pose a more realistic challenge.

In addition, proper validation is essential, and replication is a crucial piece of the validation process. As Liu et al[1] note, in machine learning studies, it remains critical to know whether the model has been validated across new clinical settings. Many of the most important challenges in machine learning are related to various forms of overfitting in which a model explains a training data set perfectly but fails to generalize. Showing a model performs well in another patient cohort in the same health system is good; showing that it performs well in an entirely different setting is far better. Such replication is the beginning, not the end, of a long process for validation and dissemination—one that draws on decades of lessons from work on developing diagnostics.

While much of the guidance in the article by Liu et al[1] will be familiar, a few key considerations bear particular emphasis in the context of machine learning applications to medicine. For example, the authors note that more complex machine learning systems are often pretrained on one data set (eg, public images on the internet of places and things) and then refit to another task (eg, retinal images). The kinds of bias introduced by such procedures is not well understood. For example, it seems likely that interpreting ophthalmologic images requires additional features beyond those needed to distinguish major categories, such as with images in general. In this case, the trained model may be systematically failing on those elements specific to opthalmology—that is, requiring features not present in general internet images while performing well overall. Such failures may be particularly concerning if they result in the model performing more poorly for specific types of patients.

The preceding example raises a larger point: because machine learning methods are myriad, in a state of rapid development, and less familiar to most clinical readers, authors of articles using machine learning must make their underlying assumptions, model properties, optimization strategies, and limitations explicit in the article. The example of transfer learning reusing a previously trained model is just one way in which properties are implicitly introduced; another is how regularization, a form of smoothing, is performed—smoothing different parameters can have different effects on the final behavior and performance the model. The predictions made by an algorithm may or may not be robust to even tiny changes to the input (eg, how differences in an image that are nondiscernible to the human eye may cause an algorithm to change its predictions).[5] Because these failure modes may not be expected, it is essential that the authors of articles reporting on machine learning point out what the failure modes of their algorithmic approaches might be. An acknowledgment of limitations should make readers more rather than less confident in clinical application. The clear parallel is precision medicine—recognizing that a medication does not work well for a subgroup of patients should only increase confidence in its use elsewhere.

The publication of machine learning work in clinical settings also requires sophistication from editors and reviewers. The onus should not be on the average clinical reader to vet the internal aspects of a machine learning inference procedure. Rather, it should be the role of the reviewers and editors to ensure that these important but highly technical aspects of the work are done correctly and presented in an understandable manner. Content reviewers may credit authors for thinking outside the box, but statistical reviewers have an obligation to think carefully about what is inside the box.

Beyond the list provided by Liu et al,[1] additional considerations that will already be familiar to readers of the diagnostic testing literature merit particular attention.

The first is the importance of subgroup analysis. As often noted, due to their complexity, machine learning models are prone to systematic errors. Readers should look for subgroup analyses of error rates across different demographics of interest and analyses showing where the greatest number of errors occur. In genomics, the problem of populations being left behind by risk modeling has gained increasing attention. The same risk applies to artificial intelligence, such as in models trained on populations that are not representative of broader communities, with the added challenge that biases in machine learing models may not be as apparent. Performance in one group or another may also reflect specifics of training parameters rather than failure to generalize. Recent recognition that machine learning in health care can both reflect and reinforce racial bias only reinforces the need to explicitly consider performance in patient subgroups in model development and deployment.[6]

The second consideration is that many machine learning studies, like many epidemiological studies, use very large cohorts. When the validation set is large, variances around effect sizes and the corresponding *P* values will be small. However, a difference that is statistically significant may not be practically meaningful. Moreover, while it may seem reasonable to conclude from a statistical perspective that the combination of flexible predictors and large training sets will result in models with both low bias and low variance, neither flexible predictors, large training sets, nor large validation cohorts provide protection from biases that come from optimization choices (eg, pretraining), choices around data processing (eg, handling missingness), or differences between the validation cohort and the populations of interest.

Liu and colleagues[1] elected to focus on applications of machine learning for diagnosis and primarily on articles addressing application for imaging. While the work in this area is likely to be the first in which machine learning could have a large effect in medicine, it is also arguably one of the most straightforward applications: in imaging studies, images arrive, are annotated, and are assessed. However, machine learning is also being applied to many other areas, such as prediction of prognosis as a means of stratifying treatment.[7-9] Especially in settings involving interventions and prognosis, it is essential that readers seek and authors provide discussions of the extent to which predictors may actually represent proxies for (unmeasured) severity that may be specific to a particular health system or setting. These circumstances do not necessarily undermine the usefulness of a model, but they should raise concern for generalizability.

For example, suppose a machine learning algorithm uses a large number of procedure and diagnostic codes as input. A complex machine learning algorithm can internally learn that the timing of various measurements (ie, a property that comes from a clinician's perspective on severity) is indicative of a certain prognosis or treatment effect. However, this learning is confounded by indication and may have modest value for prospective use: if another clinic has a different standard of care, or the original clinic changes its standard of care based on the algorithm, that algorithm will no longer provide accurate predictions. When models and inputs are simple, these errors are relatively easy to correct: model developers can make sure that the model only receives features that are properties of the patient, rather than of their care, and that the model uses those properties in a sensible manner. Machine learning algorithms help to avoid tedious manual feature engineering, that is, creating patient characteristics by hand. But, this aspect means that it may be difficult to notice when algorithms make errors that introduce confounding. Thus, even in complex models, peering inside the black box by attempting to understand the features driving predictions is important, and readers should be skeptical of any work that does not provide such justification.

With this *JAMA* how-to guide, readers should be no more intimidated by artificial intelligence than by other emerging technologies; they need not know how it works to evaluate whether it works. Results from machine learning studies are conceptually similar to results found in any other way; promising results on an independent validation set, with careful consideration of subgroup results, is just the start of a long process toward replication, prospective validation, and eventual adoption. Artificial intelligence is no more magic than logistic regression, even if it sometimes yields better results. It is necessary to use the same care in taking guidance from these sources as from their predecessors.

ARTICLE INFORMATION

Author Affiliations: Harvard University, Boston, Massachusetts (Doshi-Velez); Massachusetts General Hospital, Boston (Perlis).

Corresponding Author: Roy H. Perlis, MD, MSc, Massachusetts General Hospital, 185 Cambridge St, Sixth Floor, Boston, MA 02114 (rperlis@mgh.harvard.edu).

Conflict of Interest Disclosures: Dr Doshi-Velez reports consulting for Davita Kidney Care. Dr Perlis reports receipt of personal fees (scientific advisor services unrelated to artificial intelligence) from Psy Therapeutics, Outermost Therapeutics, Takeda, Genomind, RID Ventures, and Burrage Capital. No other disclosures were reported.

REFERENCES

1. Liu Y, Chen P-HC, Krause J, Peng L. How to read articles that use machine learning: Users' Guides to the Medical Literature. *JAMA*. doi:10.1001/jama.2019.16489

2. Nierenberg AA, Feinstein AR. How to evaluate a diagnostic marker test. *JAMA*. 1988;259(11):1699-1702.

3. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316(22):2402-2410.

4. Quinn JA, Nakasi R, Mugagga PKB, Byanyima P, Lubega W, Andama A. Deep convolutional neural networks for microscopy-based point of care diagnostics. Paper presented at: Proceedings of Machine Learning Research, Machine Learning for Healthcare Conference; August 19-20, 2016; Los Angeles, CA. Volume 56.

5. Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam AL, Kohane IS. Adversarial attacks on medical machine learning. *Science*. 2019;363(6433):1287-1289. doi:10.1126/science.aaw4399

6. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019; 366(6464):447-453. doi:10.1126/science.aax2342

7. Pineau J, Guez A, Vincent R, Panuccio G, Avoli M. Treating epilepsy via adaptive neurostimulation: a reinforcement learning approach. *Int J Neural Syst*. 2009;19(4):227-240.

8. Miotto R, Li L, Kidd BA, Dudley JT. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Sci Rep*. 2016;6(1):26094. doi:10.1038/srep26094

9. Randomised trial of intravenous streptokinase, oral aspirin, both, or neither among 17 187 cases of suspected acute myocardial infarction: ISIS-2. ISIS-2 (Second International Study of Infarct Survival) Collaborative Group. *Lancet*. 1988;2(8607):349-360.