VIEWPOINT

Machine Learning and Statistics in Clinical Research Articles—Moving Past the False Dichotomy

Samuel G. Finlayson, MD, PhD
Department of Pediatrics, Seattle Children's Hospital, Seattle, Washington; and Department of Genetics, University of Washington, Seattle.

Andrew L. Beam, PhD
Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts.

Maarten van Smeden, PhD
Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, the Netherlands.

Corresponding Author: Samuel G. Finlayson, MD, PhD, Department Pediatrics, Seattle Children's Hospital, 4800 Sand Point Way NE, OC.7.830, Seattle, WA 98105 (sgfin@uw.edu).

Medical artificial intelligence (AI) and machine learning have progressed rapidly over the past decade, yielding many new products that clinicians must increasingly learn to integrate into clinical practice.[1] A common question is, how do AI and machine learning relate to more familiar work from medical statistics?

Historical Context

In the summer of 1956, a group of computer scientists gathered at Dartmouth for a 2-month workshop to discuss what organizer John McCarthy termed *artificial intelligence*: "the science and engineering of making intelligent machines."[2] From the outset, AI attracted researchers from diverse backgrounds including neuroscience, telecommunications, and formal logic. The field was defined not by any specific methodologic approach but rather by the shared goal of enabling computers to solve new tasks.[3] Machine learning is the subfield involving a data-driven approach to AI and received its name from Dartmouth workshop attendee Arthur Samuel, who is credited as coining *machine learning* while discussing his work at IBM building a computer that plays checkers.[4] The core premise of machine learning is that a feasible path toward an intelligent computer is to build a learning computer—a machine that improves from experience and exposure to data.

Given this goal of learning from data, the field of machine learning was destined to collide with another field that came of age in the 20th century—statistics, the discipline of collecting, analyzing, and drawing conclusions from data. Like other data-centric fields such as econometrics, machine learning depends directly on statistics. Machine learning is atypical, however, in that its primary aim is not generally to generate human insights per se but rather to use analytic methods as a core component of computer systems that perform specific tasks. As researcher Tom Mitchell wrote, "The defining question for machine learning builds on both [that of computer science and statistics], but it is a distinct question."[5] Following this reasoning, discussing machine learning as a strict alternative to statistics, or vice versa, is in most cases a category error tantamount to asking if an automobile is an alternative to its engine.

Over the past half century, statisticians and computer scientists have developed a broad phylogeny of analytic methods from simple linear models to deep neural networks; the best choice among these tools is situational. Given the focus on enhancing computer performance, the practice of machine learning often favors analytic methods with high capacity to encode complex relationships among variables even if the identified patterns are harder to summarize to humans. This has led to an association of specific methods (eg, random forests, support vector machines, and neural networks) with machine learning even though many such methods were developed by statisticians and have heavily influenced their field.[6] However, the use of complex analytic models is neither necessary nor sufficient for machine learning. Indeed, many enterprise machine learning systems such as email spam filters have entailed simple statistical models deployed on a large scale. Functional machine learning systems also require the integration of analytic models into software and/or physical devices, human user interface and workflow considerations, and monitoring the resultant feedback loop as users and machine learning systems affect each others' behavior.

Moving Beyond a False Dichotomy

An unfortunate trend has emerged in recent years of emphasizing a false dichotomy between statistics and machine learning, with the latter framed not as an approach to building learning computers but rather as a specific collection of data analytic models serving as a drop-in alternative to classical statistics. This betrays a limited understanding of machine learning and its history, as machine learning was codeveloped with and is inseparable from modern statistics.[6]

We are concerned that the false statistics–machine learning dichotomy has direct negative effects on medical research. For example, the dichotomy enables using specific analytic methods (eg, random forests) to brand an analysis as machine learning, which in turn may be conflated with innovation or technical sophistication; this incentivizes some authors or reviewers to favor so-called machine learning methods even if they are not best suited for the analysis at hand. The dichotomization also blurs the wide variety of methods within each purported category, an appreciation of which is crucial to their use and evaluation. For example, consider 4 prediction models developed for the same application: a simple linear regression model, a large regression model with many polynomials and interaction terms, a small neural network with one hidden layer, and a 100-billion parameter neural network. Lumping the first 2 as statistics and the second 2 as machine learning would mask the many practical similarities between the second and third models, downplay the many unique properties of the first and fourth models relative to the others, and provide no insight into the models' intended use.

Finally, defining these fields in terms of the aforementioned false dichotomy fundamentally misses the core value proposition of machine learning research. The great promise of machine learning, especially in the past decade, has been not simply a marginal increase in accuracy when performing a classical statistical analysis on

some data set but rather the creation of computer systems that can solve entirely new sets of tasks that were previously infeasible. For example, machine learning image-processing systems have recently been approved by the US Food and Drug Administration for fully autonomous diagnosis of diabetic retinopathy and other diseases.[1] More recent work has shown that machine learning systems can generate high-resolution images on the basis of a text description and answer complex questions posed in ordinary language. Machine learning systems, when thoughtfully deployed, offer the opportunity to rethink how long-standing health care problems are framed and how clinical workflows are implemented.

Toward Clinically Useful Prediction Models

One domain in which the work of machine learning researchers and medical statisticians have increasingly coalesced is clinical prediction models, the vast majority of which still fail to be successfully implemented into clinical workflows.[7] We therefore highlight some key considerations in building and evaluating clinical prediction models.

Analytic methods for clinical prediction must be chosen based on sound understanding of both the data on which the model will operate and the circumstances of its intended use. Large, flexible models such as deep neural networks usually thrive when large corpora of data are available, data have rich internal structure (eg, imaging), and a model's use does not require a clear understanding of how the model's inputs relate to its output. Simpler methods, such as generalized linear models, may be appropriate when the goal is to interrogate the role of specific predictive factors in a clinical outcome and/or when well-calibrated risk predictions are required.[8] In all cases, researchers reporting prediction models should explain their modeling choices in terms of specific characteristics of the prediction task.

Meticulous study design is paramount when developing and reporting prediction models, especially those using advanced methods, because the prospective utility of any prediction model requires generalization to future clinical use often in the face of dynamic clinical environments.[9] If a developer anticipates a model being incorporated into a clinical workflow, they should also consider user factors, such as usability and automation bias. In short, the development of clinical prediction models must not end with fitting a model and reporting model accuracy but must also include rigorous validation, assessment of effects on patient outcomes, and once implemented, a good strategy for monitoring quality.

The advent of modern computing, coupled with a growing array of analytic methods, has opened great possibilities both for generating statistical insights and for designing useful automated computer systems. To unlock these possibilities, we must look beyond buzzwords and focus on the identification of key clinical tasks and the principled development and rigorous evaluation of well-matched methods.

ARTICLE INFORMATION

Published Online: March 20, 2023. doi:10.1001/jamapediatrics.2023.0034

Conflict of Interest Disclosures: Dr Beam reported receiving personal fees from Generate Biomedicines outside the submitted work. No other disclosures were reported.

REFERENCES

1. US Food and Drug Administration. Artificial intelligence and machine learning (AI/ML)-enabled medical devices. Accessed November 26, 2022. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices

2. McCarthy J. What is artificial intelligence? Published November 12, 2007. Accessed February 13, 2023. http://www-formal.stanford.edu/jmc/whatisai.pdf

3. McCarthy J. [Review of] Bloomfield B ed. The question of artificial intelligence. *Ann Hist Comput*. 1988;10:227.

4. Samuel AL. Some studies in machine learning using the game of checkers. *IBM J Res Develop*. 1959;3(3):210-229. doi:10.1147/rd.33.0210

5. Mitchell TM. The discipline of machine learning. Carnegie Mellon University. 2006. Accessed February 13, 2023. http://www.cs.cmu.edu/~tom/pubs/MachineLearning.pdf

6. Breiman L. Statistical modeling: the two cultures (with comments and a rejoinder by the author). *SSO Schweiz Monatsschr Zahnheilkd*. 2001;16(3):199-231. doi:10.1214/ss/1009213726

7. van Royen FS, Moons KGM, Geersing GJ, van Smeden M. Developing, validating, updating and judging the impact of prognostic models for respiratory diseases. *Eur Respir J*. 2022;60(3):2200250. doi:10.1183/13993003.00250-2022

8. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW; Topic Group "Evaluating diagnostic tests and prediction models" of the STRATOS initiative. Calibration: the Achilles heel of predictive analytics. *BMC Med*. 2019;17(1):230. doi:10.1186/s12916-019-1466-7

9. Finlayson SG, Subbaswamy A, Singh K, et al. The clinician and dataset shift in artificial intelligence. *N Engl J Med*. 2021;385(3):283-286. doi:10.1056/NEJMc2104626