Editorial

Machine Learning in Aging Research

Michelle C. Odden, PhD^{1,*} and David Melzer, MBBCh, PhD^{2,3}¹Department of Health Research and Policy, Stanford University, California. ²Epidemiology and Public Health Group, University of Exeter Medical School, UK. ³UConn Center on Aging, University of Connecticut, Farmington.*Address correspondence to: Michelle C. Odden, PhD, Department of Health Research and Policy, Stanford University, 259 Campus Drive, Stanford, CA, 94305. E-mail: mmodden@stanford.edu

Machine learning has revolutionized the technology sector with day-to-day applications ranging from prompting our choice of movies to check processing at ATMs. With the U.S. Food and Drug Administration approval of a retinal image processor (1), interest in health care applications is high. This raises the question of whether machine learning will be a powerful tool for enhancing gerontological research. In this issue, Wallace and colleagues (2) report its application to predict mortality from sleep measures. How should we evaluate this and future outputs of machine learning research?

Machine learning includes a spectrum of methods, from human specification of algorithms to fully machine-guided analysis (3). Those who have run a backwards-selection regression procedure have implemented a simple machine learning, albeit with a single procedure: *remove largest p-value variable*, repeat until *all variables have $p < .05$* . Machine learning includes supervised and unsupervised learning. In *supervised learning*, the goal is to predict an outcome (Y), given inputs (X); for example, identifying a score predicting 10-year coronary artery disease risks. In *unsupervised learning*, the goal is to find structure or patterns in inputs (X), for example, identifying which metabolites cluster together. To ensure that patterns observed are not specific to the data sampled, *cross-validation* requires a model development dataset (training data) and a separate performance evaluation dataset (test data). The test data can be partitioned from the same sample, or derived from an external dataset. Of course, performance in external datasets from the intended target population is more convincing.

In this issue, Wallace and colleagues apply a random forest machine learning to rank the 47 sleep characteristics for mortality prediction (2). Here, machine learning can conduct an unbiased ranking of the available predictors, useful as the authors had little prior knowledge. The major strength of machine learning is combining data in nonlinear and interactive ways. Humans could do this given enough time, but machine learning is far quicker. Time in bed, hours spent napping, and wake-up time emerged as most predictive. A subsample of the study population was used to measure prediction error.

When used as research rather than technical tools, machine learning results should be interpreted with caution. Beyond the technical issues are the many perennial research questions for health studies. Perhaps most pressing is whether the data used capture the likely underlying mechanisms (or at least stable proxies) and were sufficiently free of confounding and measurement error to produce answers that might ultimately help guide effective interventions? As always, good framing of research questions and good data will result in useful answers, whether using machine learning or not. Of course, the opposite is also true: limited research questions in limited data will produce limited answers, irrespective of analytic methods. The accompanying paper is presented as a prediction model, and thus future research is needed to evaluate whether the findings generalize to other populations and if the sleep measures have any mechanistic interpretation.

Machine learning is often used to improve predictive performance of models (supervised learning). In this context, the prediction method is essentially a clinical test, and should be evaluated as such, with evidence of performance for a specified purpose and target population, preferably linked to improved outcomes. Machine learning can also determine whether a condition is present, based on data provided. As mentioned, machine learning successfully diagnoses diabetic retinopathy from retinal photographs, with a c-statistics exceeding 0.99 (4). This application has great potential to save time and resources. It also has the useful feature that clinicians can readily judge the results by looking at the photographs themselves, that is, it is transparent. Prediction models for medium or longer-term outcomes seldom have this safety feature.

Unsupervised machine learnings can consolidate large sets of correlated variables into clusters or patterns: examples include social vulnerability, MRI findings, inflammatory markers, and metabolic measures (5–8). However, this approach can introduce bias and the tendency to “see” patterns reinforcing prior beliefs.

The major hazard in the machine learning and “Big Data” era is the tendency to trust high volumes of data and machine learning to produce more accurate estimates than standard approaches.

High-volume health data are often obtained from routine care settings, where recording is for nonresearch purposes with known (eg, reimbursement rules) and unknown biases incorporated into data gathering. This often results in informative patterns of missing data, increased measurement error, and unmeasured confounding (9). There are numerous examples of human biases in data subsequently encoded in statistical models (10,11). Alarming, COMPAS, a statistical risk assessment software predicting recidivism in the justice system, was more likely to incorrectly flag black defendants as future criminals, mislabeling them more than twice as often as white defendants (12). Another high-profile failure was Google Flu Trends, which aimed to identify geographic trends in flu epidemics from search patterns. The initial version performed poorly because during the learning phase, the algorithm identified search terms temporally related to flu (ie, high school basketball), but not causally. In other words, the algorithm was confounded by seasonality (13).

As researchers who have incorporated machine learning into their work, the main critique we hear from colleagues is a mistrust of “black-box” methods. This may be in part because machine learning methods were not taught in most epidemiology and clinical research curricula. However, this is changing rapidly, especially with the explosion of data science programs (14). Another issue is that *ensemble* methods, or methods that combine estimates across multiple algorithms have no simple functional form that can be written and require a certain amount of abstract understanding from the reader. Examples of ensemble methods include the random forest algorithm used in the accompanying investigation and the SuperLearner, which is an “algorithm of algorithms” and reports a weighted average of models to optimize performance (15). It seems likely that opaque methods that cannot be easily checked may find few applications in the justifiably risk averse health care arena.

Lack of reproducibility in machine learning studies has led some to proclaim that science is facing a “crisis” (16). Substantial tensions have emerged between statisticians and computer scientists on the role of machine learning in statistical analysis (17). Roots of these tensions include that many machine learning methods were developed for prediction in commercial contexts, such as the Netflix Challenge where teams competed for a \$1 million prize by predicting Netflix user movie selections. This is in sharp contrast to research aimed at identifying causal relationships or fundamental features that might be good targets for intervention. It also contrasts with prediction of health outcomes, which ultimately aim to inform interventions, where high and stable predictive performance using transparent and auditable methods are likely to be favored.

Clearly, standards for evaluating machine learning research reports are needed. Machine learning is often applied to observational data, including nonresearch data. Most of the perennial challenges of using such data in health research will remain, irrespective of whether it is Big Data or not, and irrespective of analytic methods. Assessments of machine learning reports should also aim to distinguish technical from research applications. In the meantime, it would be wise to treat machine learning research results as hypothesis generating, pending independent replications with well framed aims, high-quality data, transparent analyses, and clear accounting for confounders. However, thoughtful application of machine learning offers the exciting prospect of expanding our capacity to learn from data, which is ultimately likely to accelerate discovery and prediction in gerontology.

Acknowledgements

We would like to thank Michael T. M. Baiocchi, Ph.D. and Julia F. Simard, Sc.D. for their insightful discussion.

References

1. FDA permits marketing of artificial intelligence-based device to detect certain diabetes-related eye problems [press release]. 2018. <https://www.fda.gov/newsevents/newsroom/pressannouncements/ucm604357.htm>. Accessed March 4, 2019.
2. Wallace ML BD, Redline S, Stone K, Ensrud K, Leng Y, Ancoli-Israel S, Hall MH. Multidimensional sleep and mortality in older adults: a machine-learning comparison with other risk factors. *J Gerontol A Biol Sci Med Sci*. 2019;74:1903–1909. doi:10.1093/gerona/glz044
3. Beam AL, Kohane IS. Big data and machine learning in health care. *JAMA*. 2018;319:1317–1318. doi:10.1001/jama.2017.18391
4. 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;316:2402–2410. doi:10.1001/jama.2016.17216
5. Hames E, Pandya N, Tewary S, Stoler J, Emrich CT. A GIS approach to identifying socially and medically vulnerable older adult populations in South Florida. *Gerontologist*. 2016;57:1133–1141. doi:10.1093/geront/gnw106
6. Longstreth WT, Jr, Diehr P, Manolio TA, et al. Cluster analysis and patterns of findings on cranial magnetic resonance imaging of the elderly: the cardiovascular health study. *Arch. Neurol*. 2001;58:635–640. doi:10.1001/archneur.58.4.635
7. Sakkinen PA, Wahl P, Cushman M, Lewis MR, Tracy RP. Clustering of procoagulation, inflammation, and fibrinolysis variables with metabolic factors in insulin resistance syndrome. *Am J Epidemiol*. 2000;152:897–907. doi:10.1093/aje/152.10.897
8. Mukamal KJ, Siscovick DS, de Boer IH, et al. Metabolic clusters and outcomes in older adults: the cardiovascular health study. *J Am Geriatr Soc*. 2018;66:289–296. doi:10.1111/jgs.15205
9. National Academies of Sciences, Engineering and Medicine. *Refining the Concept of Scientific Inference When Working with Big Data: Proceedings of a Workshop*. Washington, DC: The National Academies Press; 2017. doi:10.17226/24654
10. Garcia M. Racist in the machine: the disturbing implications of algorithmic bias. *World Policy Journal*. 2016;33:111–117. doi:10.1215/07402775-3813015
11. O’Neil C. *Weapons of math destruction: How big data increases inequality and threatens democracy*. 1st ed. New York: Crown Publishers; 2016.
12. Angwin J, Larson J, Mattu S, Kirchner L. Machine bias: There’s software used across the country to predict future criminals and it’s biased against blacks. 2016. <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>. Accessed March 4, 2019.
13. Lazer D, Kennedy R, King G, Vespignani A. The parable of google flu: traps in big data analysis. *Science*. 2014;343(6176):1203–1205. doi:10.1126/science.1248506
14. Topco G. Connecting Data Science to ‘Almost Every Domain of Inquiry’. 2018; <https://www.insidehighered.com/news/2018/11/02/big-data-ai-prompt-major-expansions-uc-berkeley-and-mit>. Accessed March 4, 2019.
15. Rose S. Mortality risk score prediction in an elderly population using machine learning. *Am J Epidemiol*. 2013;177:443–452. doi:10.1093/aje/kws241
16. Ghosh P. AAAS: Machine learning ‘causing science crisis’. 2019. <https://www.bbc.com/news/science-environment-47267081>. Accessed March 4, 2019.
17. Donoho D. 50 years of data science. *J. Comput. Graph. Stat*. 2017;26:745–766. doi:10.1080/10618600.2017.1384734