EDITORIAL

Incremental Benefits of Machine Learning—When Do We Need a Better Mousetrap?

Matthew M. Engelhard, MD, PhD; Ann Marie Navar, MD, PhD; Michael J. Pencina, PhD

+ Related article

It is indisputable that clinical medicine has entered the age of big data. Newer, better prediction methods, such as neural networks, random forests, and other algorithms, often categorized as *machine learning* (ML), can probe large-scale clinical data sets to discover predictive features unavailable to their more traditional counterparts. However, when Khera et al[1] pit 3 of these algorithms against the most standard generalized linear model, logistic regression, to predict death after acute myocardial infarction, none of the ML algorithms emerge as a clear winner. Two of the 3 ML algorithms improved discrimination by a slim margin and yielded "more precise calibration across the risk spectrum."[1] However, these improvements are unlikely to be clinically meaningful, and it's unclear whether they would be sufficient to justify the corresponding loss of interpretability. Furthermore, 1 ML approach (a neural network) performed worse than logistic regression. The data set in question is undeniably big, at least in sample size. Khera et al[1] draw on an American College of Cardiology registry that contains more than 750 000 records; therefore, at first glance, it appears that the promise of using ML to harness big data is not being realized. What can explain this disconnect, and does it suggest that ML is more hype than substance?

This is not the first time that ML has offered only modest or no improvement of traditional regression models in clinical medicine. For every study in which it yielded superior results, there are others in which the gains disappoint. Indeed, the balance of evidence suggests that for typical clinical prediction tasks, wherein predictions are based on a modest number of clinical variables, ML algorithms are on par with logistic regression.[1,2] Electronic health records (EHRs) contain thousands of potential predictor variables, so we might expect ML algorithms to fare better; yet to date traditional approaches still appear to hold their own. For the prediction of death, readmission, and length of stay from the EHR, for example, a logistic regression model with access to the same predictors was competitive with deep learning, with differences between them often falling within the margin of error.[3] In principle, ML models are more flexible than logistic regression and other generalized linear models: they can more accurately approximate the true relationship between predictors and outcome when data are plentiful. In practice, however, logistic regression is a tough baseline to beat.

Although ML has been underwhelming for many prediction tasks, it has been a profound advance for others. For example, in medical image processing, prediction models have reached levels of performance unimaginable only a decade ago, largely because of the success and popularization of deep convolutional neural networks. Machine learning had exceeded human performance when classifying everyday images in 2015, and only 2 years later, a convolutional neural network model was found to perform as well as experienced dermatologists when identifying skin lesions.[4] Similar findings have been reported across a range of imaging-intensive medical specialties, including ophthalmology, pathology, and radiology. Within the past year, an ML algorithm was trained to read mammograms more effectively than radiologists.[5] For these problems, in stark contrast to the prediction tasks previously discussed, the role of ML is not to make predictions better but rather to allow prediction at all. Thus, there is no comparison to logistic regression because linear models are entirely inadequate for image processing. Medical images, and indeed all natural images, are composed of millions of pixels interconnected in a rich, complex spatial structure, and any given pixel has little to no meaning in isolation. Consequently, ML methods are required to decipher them.

Although image processing might be the best known success of ML thus far, it is far from the only one. Machine learning is commonplace in digital health applications, in which wearable devices continuously generate high-dimensional, highly structured time series measuring physical activity and a range of physiologic parameters. Much like pixels in an image, individual samples from a wearable device are part of a rich temporal structure, and as a result, deep neural networks are state of the art for many digital health prediction tasks, such as human activity recognition. Similarly, deep neural network models are often the accepted method for processing physiologic time series recorded in the laboratory or clinic, a notable example being electroencephalography.[6] In natural language processing, in which ML models such as the support vector machine have long outperformed traditional models, a milestone has been reached: deep learning models now exceed human-level performance in a variety of tasks, including question answering. These models are the current state of the art for processing biomedical and clinical text[7] and have begun to appear in the clinical literature. In addition, an increasing number of clinical prediction tasks are based on a fusion of modalities, such as images combined with EHR variables,[8] thus necessitating a ML approach. In each of these examples, as with image processing, ML is not just an incremental improvement but is often the difference between an algorithm that is useful and one that is not.

What do these success stories have in common? In the cases in which ML has been most impactful, the data are high dimensional, highly structured, and difficult to summarize without substantial loss of information. In a word, they are complex. In the cases in which it has been least impactful, there tend to be fewer predictors, and unlike in images or text, individual predictors more often correspond to distinct measurements or attributes. The characteristics of the data at hand are therefore central when developing a prediction model; it is important to choose a modeling approach that is well matched to these characteristics. To aid in the initial steps of this process, we propose the following rule of thumb: as the complexity of the underlying data increases, so too does the value of ML (**Figure**).

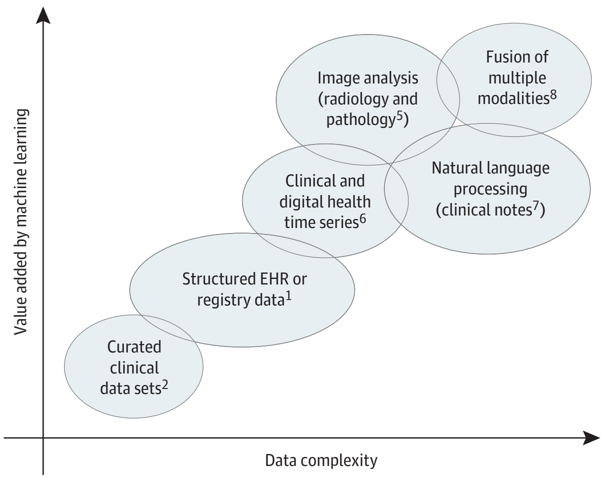
When working with images, text, or time series, ML is almost sure to add value, whereas when working with a fewer, weakly correlated clinical variables, logistic regression is likely to do just as well. In the substantial gray area between these extremes, judgment and experimentation are required. The work by Khera et al[1] falls in this gray area and sheds light on the potential that ML can offer: when correctly applied, it might lead to more meaningful gains in calibration than discrimination. This is an important finding, because the role of calibration is increasingly recognized as key for unbiased clinical decision-making, especially when threshold-based classification rules are used. The *correctly applied* caveat is also important; unfortunately, many developers of ML models treat calibration as an afterthought.

The value in complexity principle illustrated in the Figure is likely an oversimplification and should be reevaluated as new algorithms are developed and data resources continue to expand. Research into ML is progressing at a rapid pace, and several areas in current development are directly relevant to health care. Particularly relevant might be the ongoing work in representation learning for EHR data elements, which aims to numerically encode relationships between medical codes and/or concepts and may continue to increase the value of deep learning for EHR-based prediction tasks.[9] Despite these caveats, we believe that this rule of thumb might be useful in guiding the selection of mathematical models for the task at hand and setting realistic expectations for possible improved performance.

Provided the model evaluation is conducted properly, there is little harm in exploring whether an artificial neural network—or any other ML algorithm—may improve on logistic regression for a given clinical prediction task. However, moving away from the generalized linear model is not without cost. Although feature attribution methods for ML models continue to improve, linear models remain more interpretable because they are simple and familiar to most practitioners and researchers. Complex models also tend to be more complicated to implement, and they are undoubtedly more costly to develop, partly because development requires different skills and software compared with the more traditional approach. As the results presented by Khera et al[1] illustrate, the generalized linear model is powerful, and only rarely is there a price—a substantial loss of performance—for choosing it. When developing a prediction model, we should choose the simplest tool that will do the job.[10] By honing our intuitions about the likely value added by ML, we can maximize our efforts and sacrifice the simplicity and interpretability of the linear model only when necessary.

Recent feats of ML in clinical medicine have seized our collective attention, and more are sure to follow. As medical professionals, we should continue building familiarity with these technologies and embrace them when benefits are likely to outweigh the costs, including when working with complex data. However, we must also recognize that for many clinical prediction tasks, the simpler approach—the generalized linear model—may be all that we need.

Figure. Examples of the Value Added by Machine Learning When Applied to Data With Increasing Complexity


Value added by machine learning
Data complexity
Image analysis (radiology and pathology[5])
Fusion of multiple modalities[8]
Clinical and digital health time series[6]
Natural language processing (clinical notes[7])
Structured EHR or registry data[1]
Curated clinical data sets[2]

As the complexity of the underlying data increases, so too does the probable value added by machine learning. EHR indicates electronic health record.

ARTICLE INFORMATION

Author Affiliations: Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, North Carolina (Engelhard, Pencina); Department of Internal Medicine, UT Southwestern Medical Center, Dallas, Texas (Navar); Department of Population and Data Sciences, UT Southwestern Medical Center, Dallas, Texas (Navar); Associate Editor, *JAMA Cardiology* (Navar); Duke Clinical Research Institute, Durham, North Carolina (Pencina); Deputy Editor for Statistics, *JAMA Cardiology* (Pencina).

Corresponding Author: Michael J. Pencina, PhD, Duke Clinical Research Institute, Duke University, 200 Trent Dr, Room M144, Davison Building, Durham, NC 27710 (michael.pencina@duke.edu).

Published Online: March 10, 2021. doi:10.1001/jamacardio.2021.0139

Conflict of Interest Disclosures: Dr Navar reported receiving personal fees from Cerner Corporation and *JAMA Cardiology* outside the submitted work. Dr Pencina reported receiving grants from Sanofi/Regeneron and Amgen (to Duke University) and personal fees from Boehringer Ingelheim outside the submitted work. No other disclosures were reported.

REFERENCES

1. Khera R, Haimovich J, Hurley N, et al. Use of machine learning models to predict death after acute myocardial infarction. *JAMA Cardiol*. Published online March 10, 2021. doi:10.1001/jamacardio.2021.0122

2. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol*. 2019;110:12-22. doi:10.1016/j.jclinepi.2019.02.004

3. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. *NPJ Digit Med*. 2018;1(1):18. doi:10.1038/s41746-018-0029-1

4. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542(7639):115-118. doi:10.1038/nature21056

5. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. *Nature*. 2020;577(7788):89-94. doi:10.1038/s41586-019-1799-6

6. Craik A, He Y, Contreras-Vidal JL. Deep learning for electroencephalogram (EEG) classification tasks: a review. *J Neural Eng*. 2019;16(3):031001. doi:10.1088/1741-2552/ab0ab5

7. Huang K, Altosaar J, Ranganath R. ClinicalBERT: modeling clinical notes and predicting hospital readmission. *arXiv*. Preprint posted online November 28, 2020. doi:10.1145/1122445.1122456

8. Huang S-C, Pareek A, Seyyedi S, Banerjee I, Lungren MP. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *NPJ Digit Med*. 2020;3(1):136. doi:10.1038/s41746-020-00341-z

9. Choi E, Xiao C, Stewart WF, Sun J. MiME: multilevel medical embedding of electronic health records for predictive healthcare. *arXiv*. Preprint posted online October 22, 2018.

10. Pencina MJ, Goldstein BA, D'Agostino RB. Prediction models: development, evaluation, and clinical application. *N Engl J Med*. 2020;382(17):1583-1586. doi:10.1056/NEJMp2000589