VIEWPOINT

Making Machine Learning Models Clinically Useful

Nigam H. Shah, MD, PhD
Stanford Center for Biomedical Informatics Research, Stanford University, Stanford, California.

Arnold Milstein, MD, MPH
School of Medicine, Stanford University, Stanford, California.

Steven C. Bagley, PhD
Stanford Center for Biomedical Informatics Research, Stanford University, Stanford, California.

Recent advances in supervised machine learning have improved diagnostic accuracy and prediction of treatment outcomes, in some cases surpassing the performance of clinicians.[1] In supervised machine learning, a mathematical function is constructed via automated analysis of training data, which consists of input features (such as retinal images) and output labels (such as the grade of macular edema). With large training data sets and minimal human guidance, a computer learns to generalize from the information contained in the training data. The result is a mathematical function, a model, that can be used to map a new record to the corresponding diagnosis, such as an image to grade macular edema. Although machine learning–based models for classification or for predicting a future health state are being developed for diverse clinical applications, evidence is lacking that deployment of these models has improved care and patient outcomes.[2]

One barrier to demonstrating such improvement is the basis used to assess the performance of a model. Current approaches gauge performance by quantifying how closely the diagnosis or prediction made by the model matches known diagnoses or health outcomes. Quantifications include sensitivity, specificity, and positive predictive value, as well as measures such as the area under the receiver operating characteristic (ROC) curve, the area under the precision-recall curve, and calibration. Because no single measurement reflects all of the desirable properties of a model, several measurements typically are reported to summarize the performance of the model.

> [R]ealizing the potential benefit of machine learning for patients in the form of better care requires rethinking how model performance during machine learning is assessed.

However, none of the measures address the purpose of the model or what matters most to patients, ie, that a classification or prediction from the model results in a favorable change in their care. Given a model's prediction, whether actions are taken and the effect those actions may have is determined by numerous factors in the health care system. These include a clinician's capacity to formulate a responsive action, weigh its risks and benefits, and execute the action, as well as the patient's adherence with the clinical action recommended. For example, there are multiple models to predict hospital readmissions, with machine learning–derived models showing higher predictive accuracy than traditional factors, such as diagnosis and demographic factors.[3] Success in reducing readmissions remains limited because of constraints such as clinician time, available staff, and limited ability to influence social determinants of health.[4]

The limitation of not considering the characteristics of the care environment in evaluating the performance of a model applies to all types of classification and prediction models and not just to models derived via machine learning. There are methods to assess whether the use of a machine learning model will be useful given the prevailing constraints in the care environment. For example, given the costs of alternative actions, their corresponding benefits, and the various measures of model performance, methods such as decision curve analysis can quantify the net benefit of using a model to guide subsequent actions.[5]

A necessary next step in evaluating the performance of classification or prediction models is to estimate the net incremental value of taking plausible alternative actions for patients. However, such analysis is usually not performed for machine learning–derived models. In the few instances that such analyses have been done, they were performed after selecting the best-performing model. A 2-step process of selecting a best model and then evaluating whether the model is useful can be misleading because in machine learning hundreds of computerized models are created from which 1 is selected during the process of learning.

For simplicity consider 2 models in the **Figure**, which shows ROC curves for 2 readmissions prediction models with the orange curve having a larger area under the ROC curve (0.79) than does the blue curve (0.74). Given estimated costs and benefits of the actions possible to try to prevent a readmission, the light blue triangle highlights the region of higher utility than what the best model allows to be achieved. The blue curve has a smaller area under the ROC curve, but taking readmission-preventing actions based on that model's predictions has higher utility because the blue curve crosses over into the light blue region. In fact, several other models could be created during the process of learning that have even higher utility, but those will never be considered because the choice of the best model is based on measures such as the area under the ROC curve. This example illustrates how a 2-step process to evaluate net benefit will fail to uncover that a model more useful than the best model, based on the area under the ROC curve, exists.

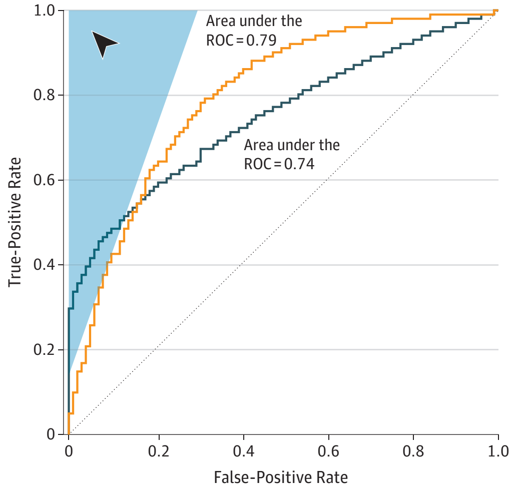
When decision analysis is conducted after selecting the best model, the constraints of the subsequent actions cannot inform the model selection that occurs during machine learning. Therefore, characteristics such as the number of actions the care team can take, the cost

Corresponding Author: Nigam H. Shah, MD, PhD, Stanford Center for Biomedical Informatics Research, Stanford University, 1265 Welch Rd, Stanford, CA 94305 (nigam@stanford.edu).

© 2019 American Medical Association. All rights reserved.

Downloaded From: https://jamanetwork.com/ on 08/08/2019

Figure. Considering Net Benefit During the Selection of the Best-Performing Model



Two receiver operating characteristic (ROC) curves are shown. Expected utility increases in the direction of the arrowhead. The hypotenuse of the light blue triangle marks the maximum utility achievable by taking readmission preventing actions based on the model represented by the orange curve. The blue ROC curve extends into the triangle, showing that actions based on this model have a higher utility, even though it has a lower area under the ROC.

and presumed efficacy of those actions, and the chance that the patient will follow the recommended action need to be considered concurrently when a computer is learning the model from the training data. Doing so will require integration of data from other information systems regarding constraints on availability of essential resources, such as personnel, space, and equipment. The availability of such data is growing. However, the use of such data during machine learning to anticipate the net incremental value of making and responding to a classification or prediction within the constraints of the care environment is lacking.

Even after the assessment of models evolves to consider constraints of the care environment, it must be clear who has responsibility for taking the action.[2] Weighing the societal, cultural, and human contexts to choose the appropriate action for the individual patient requires human judgment.[6] Human judgment is also required to assess patient preferences regarding the use of a model as well as ethical and medicolegal issues, such as unwanted diffusion of responsibility or new obligations to treat.[7] Personalizing the choice of the subsequent actions may be necessary to realize the benefit from having a patient-specific prediction. Such personalized action selection can be guided by an analysis of clinician characteristics such as available expertise and prior actions taken in similar situations.

Machine learning can identify patterns in the expanding, heterogeneous data sets to create models that accurately classify a patient's diagnosis or predict what a patient may experience in the future. However, realizing the potential benefit of machine learning for patients in the form of better care requires rethinking how model performance during machine learning is assessed. A framework for rigorously evaluating the performance of a model in the context of the subsequent actions it triggers is necessary to identify models that are clinically useful.

ARTICLE INFORMATION

Published Online: August 8, 2019. doi:10.1001/jama.2019.10306

Conflict of Interest Disclosures: Dr Shah reported being a cofounder of Cardinal Analytx, which uses machine learning to better predict and responsibly contain health care costs, while also improving quality of care. Dr Milstein reported receiving personal fees from Cardinal Analytx and Dawnlight, and Stanford University owns intellectual property arising from his prior research as a faculty member. Dr Bagley reported receiving grants from Google.

REFERENCES

1. Beam AL, Kohane IS. Big data and machine learning in health care. *JAMA*. 2018;319(13):1317-1318. doi:10.1001/jama.2017.18391

2. Emanuel EJ, Wachter RM. Artificial intelligence in health care: will the value match the hype? *JAMA*. 2019;321(23):2281-2282 . doi:10.1001/jama.2019.4914

3. Morgan DJ, Bame B, Zimand P, et al. Assessment of machine learning vs standard prediction rules for predicting hospital readmissions. *JAMA Netw Open*. 2019;2(3):e190348. doi:10.1001/jamanetworkopen.2019.0348

4. Braet A, Weltens C, Sermeus W. Effectiveness of discharge interventions from hospital to home on hospital readmissions: a systematic review. *JBI Database System Rev Implement Rep*. 2016;14(2):106-173. doi:10.11124/jbisrir-2016-2381

5. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making*. 2006;26(6):565-574. doi:10.1177/0272989X06295361

6. Verghese A, Shah NH, Harrington RA. What this computer needs is a physician: humanism and artificial intelligence. *JAMA*. 2018;319(1):19-20. doi:10.1001/jama.2017.19198

7. Char DS, Shah NH, Magnus D. Implementing machine learning in health care—addressing ethical challenges. *N Engl J Med*. 2018;378(11):981-983. doi:10.1056/NEJMp1714229

© 2019 American Medical Association. All rights reserved.

Downloaded From: https://jamanetwork.com/ on 08/08/2019