JACC: ADVANCES VOL. 2, NO. 7, 2023

EDITORIAL COMMENT

Machine Learning for Risk Prediction

Does One Size Really Fit All?*

Collin M. Stultz, MD, PhD[a,b]

The advent of generative models, like ChatGPT, has propelled machine learning to the forefront of the general public's mind. The ability to use such large language models to answer a wide range of questions from completely different domains is both impressive and concerning. Indeed, it begs the question—can a single model be "smart" enough to correctly answer any question it is posed? While models like ChatGPT perform well, on average, across a variety of different tasks, it has important failure modes that have yet to be comprehensively studied.

Machine learning models, unfortunately, are never one-size-fits-all algorithms. In reality, machine learning rests on the premise that machines can learn useful insights when given enough training data. A model is therefore not guaranteed to perform well on data that are dissimilar to the data upon which it was trained. Nevertheless, it is difficult to resist the urge to explore the boundaries of what a model has learned by testing its performance on new data that may not be obviously related to the training data. Such experiments can, in fact, provide insight into what the model has learned and help one understand how and when the model can fail.

In this issue of *JACC: Advances*, Jang et al[1] conduct a thorough and insightful study exploring how a previously developed machine learning model, MARKER-HF (Machine learning Assessment of RisK and EaRly mortality in Heart Failure), which was designed to predict mortality in patients with heart failure, performs on cohorts that have an array of different diagnoses. Firstly, the authors demonstrate that MARKER-HF retains its discriminatory ability with respect to predicting mortality in heart failure patients in a large community-based hospital registry from Kyungpook National University Hospital. More impressively, they discover that Marker-HF's discriminatory ability is largely maintained in cohorts representing patients who have no prior diagnosis of heart failure. The fact that MARKER-HF performance is good across a panoply of common disorders suggests that it may be a "one-size-fits-all" model for risk prediction across a variety of different diseases; ie, patients with recent acute coronary syndrome or a history of atrial fibrillation, chronic obstructive pulmonary disease, chronic kidney disease, diabetes mellitus, hypertension, and malignancy.

The MARKER-HF score was constructed using a machine learning method called a decision tree.[2] A decision tree is essentially a complex, "tree-like" flowchart that encapsulates a series of decisions and corresponding consequences. In the present application, given a particular patient's list of clinical characteristics, one can follow the instructions listed in this flowchart to decide whether this patient is at high risk of death or not. Now, there are some subtleties here, as the authors use "Adaptive Boosting"—a process that involves constructing many different decision trees and combining the predictions from these trees to arrive at a final prediction. The process of learning the model involves learning the optimal set of decisions for each decision tree and how to combine the predictions from each tree without overfitting the model to the training data. Model overfitting happens when the model learns to perform well on the training data, but performs poorly on examples that were not in the training set.

*Editorials published in *JACC: Advances* reflect the views of the authors and do not necessarily represent the views of *JACC: Advances* or the American College of Cardiology.

From the [a]Department of Electrical Engineering and Computer Science, and Institute for Medical Engineering & Science, MIT, Cambridge, Massachusetts, USA; and the [b]Division of Cardiology, MGH, Boston, Massachusetts, USA.

The author attests they are in compliance with human studies committees and animal welfare regulations of the author's institution and Food and Drug Administration guidelines, including patient consent where appropriate. For more information, visit the Author Center.

ISSN 2772-963X https://doi.org/10.1016/j.jacadv.2023.100552

Jang et al,[1] make a compelling case that their decision tree model is indeed performant on cohorts distinct from their training data, and, more importantly, performant on a cohort from another country. This is a very important result, the significance of which is hard to overstate. A major failure mode of many machine learning models is that they do not generalize outside of the home institution where they were trained and tested. In this regard, the authors have gone a long way to demonstrate the utility of their method across different heart failure populations. Nonetheless, while this work is a valiant effort and an important advance, like all interesting and provocative studies, it raises a number of questions that have yet to be answered.

MARKER-HF uses only 8 clinical variables to calculate a risk score: diastolic blood pressure, creatinine, blood urea nitrogen, hemoglobin, white blood cell count, platelets, albumin, and red blood cell distribution width. These variables were selected using an iterative approach that selects “the most common and discriminating subset of variables out of those available in the UCSD cohort.”[2] An examination of the 8 features used in MARKER-HF provides insight into why it performs adequately across a variety of disorders, as several of these clinical variables were already known to be independent predictors of death in different populations.[3-8] It is therefore likely that the performance of the MARKER-HF is wholly explained by the fact that the chosen features are quite powerful for discriminating between high-risk and low-risk patients; the upshot being that sophisticated machine learning methods may not be needed for this task. To be precise, a simple logistic regression model or a Cox proportional hazards model—both of which are standard statistical techniques—developed using these 8 features may perform just as well as MARKER-HF. As this comparison was not done in this paper, it remains unclear whether Adaptive Boosting Decision Trees (the method used to develop the MARKER-HF score) was needed at all.

While the areas under the receiver operator curve (AUCs) of MARKER-HF are good over a range of different diagnoses, these data are not sufficient to truly understand how to use the model in practice. All too often, machine learning practitioners (and I count myself a member of this community) rely on the AUC as our main metric of success. However, good discriminatory ability is only a necessary condition for a good model, and consequently, a good AUC does not guarantee that a model is ready for prime time. For example, while the MARKER-HF AUC for 1-year mortality predictions is 0.738, the authors report a corresponding sensitivity of 0.65. In fact, the reported sensitivity for predicting 1-year death in HF patients—the task the MARKER-HF was designed to tackle—is only 0.68. In fairness to the authors, the sensitivity is a function of the MARKER-HF threshold one uses to make a decision about whether a patient is actually at high risk or not, so this is in principle modifiable. However, the authors state that this value corresponds to the optimal threshold and the criterion used to choose it remains unclear. For outcomes such as death, it is desirable to have sensitivities that are considerably higher.

MARKER-HF is a promising tool that will likely find a welcome role in our armamentarium of risk stratification methods for heart failure patients. Whether it is truly a one-size fits all model, however, remains to be seen.

FUNDING SUPPORT AND AUTHOR DISCLOSURES

The authors has reported that he has no relationships relevant to the contents of this paper to disclose.

ADDRESS FOR CORRESPONDENCE: Dr Collin M. Stultz, MIT, Building 36-796 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA. E-mail: cmstultz@mit.edu.

REFERENCES

1. Jang SY, Park JJ, Adler E, et al. Mortality prediction in patients with or without heart failure using a machine learning model. *JACC: Adv*. 2023;2:100554.

2. Adler ED, Voors AA, Klein L, et al. Improving risk prediction in heart failure using machine learning. *Eur J Heart Fail*. 2020;22:139–147.

3. Maaravi Y, Bursztyn M, Hammerman-Rozenberg R, Cohen A, Stessman J. Moderate renal insufficiency at 70 years predicts mortality. *QJM*. 2006;99:97–102.

4. Wernly B, Lichtenauer M, Vellinga NAR, et al. Blood urea nitrogen (BUN) independently predicts mortality in critically ill patients admitted to ICU: a multicenter study. *Clin Hemorheol Microcirc*. 2018;69:123–131.

5. Kabat GC, Kim MY, Manson JE, et al. White blood cell count and total and cause-specific mortality in the Women's Health Initiative. *Am J Epidemiol*. 2017;186:63–72.

6. Tsai MT, Chen YT, Lin CH, Huang TP, Tarng DC. U-shaped mortality curve associated with platelet count among older people: a community-based cohort study. *Blood*. 2015;126:1633–1635.

7. Goldwasser P, Feldman J. Association of serum albumin and mortality risk. *J Clin Epidemiol*. 1997;50:693–703.

8. Salvagno GL, Sanchis-Gomar F, Picanza A, Lippi G. Red blood cell distribution width: a simple parameter with multiple clinical applications. *Crit Rev Clin Lab Sci*. 2015;52:86–105.

KEY WORDS decision trees, machine learning, risk stratification