MACHINE LEARNING AND READMISSION: DO WE NEED NEW METHODS TO SOLVE

OLD PROBLEMS?

To the Editor:

We read with great interest the study by Bolourani and colleagues¹ in which they developed a prediction model for readmission postesophagectomy using machine learning (ML) methodology. The authors reported improved performance as measured by the area under the receiver operating characteristic curve for several ML techniques compared with conventional logistic regression. Readmission after cardiothoracic surgery is common²; predicting readmission is notoriously challenging, and the c-statistics for models are typically between 0.6 and 0.7, unlike mortality predictions models, where c-statistics are between 0.8 and 0.9.³ Thus, the pursuit of novel techniques to improve our ability to predict and risk stratify patients for readmission after discharge is much welcomed.

With the increasing adoption of electronic patient records which, in turn, have generated large-scale data, and the advent of advances in data manipulation and computing power, we have seen the use of ML methodologies in medicine grow exponentially.⁴ However, with the introduction of novel methods, like ML, we should be fastidious about consistently reporting all the relevant results. This allows for the readership of clinicians and computer scientists to better understand the authors' workflow and comparison of various ML models. There have been several publications that provide guidance on the publication and reporting of ML models.⁵ Comprehensive and appropriate reporting allows the reader to draw their own conclusions regarding the findings of these models. This has led us to some questions regarding the model by Bolourani and colleagues.

In model development, given that ML is highly sensitive to class imbalances and since there were only 383 patients of the 2000 patients had the event of interest, we are

curious as to how the authors managed the learning imbalance. In addition, with respect to evaluation, while the authors did report confusion matrices, sensitivity, specificity, and receiver operating characteristic curves for their models, which are all measures of discrimination, we noted that calibration measurements were not reported. These measurements are crucial in medical applications, as these metrics provide us with a probability of certainty for many of the binary decisions that are made in medicine.⁵ We suspect that the confusion matrix magnitude measurement reflects calibration, but this is unclear in the manuscript. In addition, to appropriately compare the performance of ML and regression, we must be able to compare the same performance metrics; however here, the authors chose to present sensitivity and specificity for ML and sensitivity and accuracy for regression. We would like to better understand why the authors specifically chose these metrics to present for the respective models.

Finally, ML methods are frequently criticized as “black boxes” and as such, have a hard time performing clinically in circumstances in which further decision making requires an understanding of the model estimates. Furthermore, in conventional logistic regression models, beta-coefficients can be expressed as odds ratios, and a scoring system can be derived to be used at the patient's bedside⁶; how this can be applied to ML models is likely more complicated. While much “backend” work in model development is being pursued, we look forward to seeing ML prediction models be adapted for clinical use in user friendly applications (ie, “front-end” development).

Rashmi Nedadur, MD
Derrick Y. Tam, MD, PhD
Stephen E. Frenes, MD, MSc
Division of Cardiac Surgery
Schulich Heart Centre
Department of Surgery
Sunnybrook Health Sciences Centre
University of Toronto
Toronto, Ontario, Canada

References

1. Bolourani S, Tayebi MA, Diao L, Wang P, Patel V, Manetta F, et al. Using machine learning to predict early readmission following esophagectomy. *J Thorac Cardiovasc Surg.* 2021;161:1926-39.
2. Han S, Smith TS, Gunnar W. Descriptive analysis of 30-day readmission after inpatient surgery discharge in the Veterans Health Administration. *JAMA Surg.* 2014;149:1162-8.
3. Tam DY, Fang J, Tran A, Tu JV, Ko DT, Deb S, et al. A clinical risk scoring tool to predict readmission after cardiac surgery: an Ontario administrative and clinical population database study. *Can J Cardiol.* 2018;34:1655-64.
4. Handelman GS, Kok HK, Chandra RV, Razavi AH, Lee MJ, Asadi H. eDoctor: machine learning and the future of medicine. *J Intern Med.* 2018;284:603-19.
5. Chen P-HC, Liu Y, Peng L. How to develop machine learning models for healthcare. *Nat Mater.* 2019;18:410-4.

The Editor welcomes submissions for possible publication in the Letters to the Editor section that consist of commentary on an article published in the Journal or other relevant issues. Authors should: • Include no more than 500 words of text, three authors, and five references. • Type with double-spacing. • See <http://jtcvs.ctsnetjournals.org/misc/fora.shtml> for detailed submission instructions. • Submit the letter electronically via jtcvs.editorialmanager.com. Letters commenting on an article published in the JTCVS will be considered if they are received within 6 weeks of the time the article was published. Authors of the article being commented on will be given an opportunity to offer a timely response (2 weeks) to the letter. Authors of letters will be notified that the letter has been received. Unpublished letters cannot be returned.

The authors reported no conflicts of interest.

The *Journal* policy requires editors and reviewers to disclose conflicts of interest and to decline handling or reviewing manuscripts for which they may have a conflict of interest. The editors and reviewers of this article have no conflicts of interest.

6. Austin PC, Lee DS, D'Agostino RB, Fine JP. Developing points-based risk-scoring systems in the presence of competing risks. *Stat Med*. 2016;35:4056-72.

<https://doi.org/10.1016/j.jtcvs.2020.07.102>

REPLY: THE STANDARDIZATION AND AUTOMATION OF MACHINE LEARNING

FOR BIOMEDICAL DATA

Reply to the Editor:

As noted in our commentary on the work by Bolourani and colleagues,¹ the problem of class imbalance in biomedical data is ubiquitous and requires special care because of the tendency of machine learning (ML) methods to classify members of the majority class (ie, “negatives” or non-events) correctly at the expense of members of the minority class (ie, “positives” or events). As discussed, recent work on “quantile classifiers” is a promising new approach to navigate this challenging problem. It was thus with great interest that we read the commentary by Nedadur and colleagues² on the same article. In their commentary, they advocated for more comprehensive and standardized reporting of ML methods to promote easier understanding and evaluation, which was made more difficult in this analysis because of the nature of the data.

We sympathize with these comments, and as ML methods become more widely used, there will undoubtedly be more calls like this, as well as for standardized workflows for ML methods. (Indeed Nedadur and colleagues² provided such an example.) With this in mind, we would like to suggest 2 points for consideration in such discussions.

First, we believe that greater awareness of the limitations of current metrics used for evaluating performance is needed. C-statistics and areas under receiver operating characteristic (ROC) curves are routinely reported, but these have many limitations, most notably in the class-imbalanced setting. Specifically, they account for neither prevalence (ie, the frequency of the minority class) nor unequal misclassification costs (eg, the cost of incorrectly classifying an early readmission vs cost of incorrectly classifying a nonearly readmission). Precision-recall (PR) curves have been suggested as an alternative to ROC curves in

the presence of class imbalance.³ PR curves, however, do not have the desirable property of a “universal baseline.” Specifically, the baseline for all ROC curves is 0.5, whereas the baseline for a PR curve is the prevalence, which is data dependent and this can lead to confusing results if the same method is used to analyze new data with a slightly different prevalence.⁴ Importantly, both skirt the issue of selecting a threshold and thus evaluate *theoretic* capability rather than *actual* performance. Moreover, both include clinically nonsensical thresholds that can contribute more to the area than clinically desirable thresholds.⁵ As advocated by Nedadur and colleagues,² we recommend that calibration information for class probability estimates be provided, such as a “smoothed calibration curve” that compares predicted probability to observed proportions.⁶ Regardless, if ROC and PR curves are used, we recommend also reporting the G-mean, a robust and interpretable metric summarizing classification performance.

A second issue is the overemphasis on prediction performance, with many articles devoting entire sections and lengthy appendices to this, but with relatively less effort being placed on clinical insight. We concur with the call of Nedadur and colleagues² for meaningful and interpretable statements, but disagree that ML methods are “black boxes” not designed for this goal. In fact, there exist many tools that can aid in this endeavor. These include (1) variable importance, which provides a direct interpretation of the contribution of each variable to prediction of the outcome^{7,8}; and (2) partial effects (also known as marginal effects), which estimate the effect of a change in a specific predictor variable on the outcome after averaging over all other predictor variables.⁹ It should be noted that variable importance and partial effects are not available for all ML methods, and this should be an important consideration when choosing a procedure. Algorithms designed solely for prediction, such as deep learning and support vector machines, are true black boxes. These methods may utilize synthetic variables unrelated to original variables, thus rendering them unable to provide meaningful variable importance, and as they are only suitable for classification and recognition tasks, they do not provide class probability estimates and thus cannot provide partial or marginal effects. In contrast, there are ML methods that yield insight, such as random forests and gradient boosted trees. These methods work directly with the outcomes and original variables, even categorical variables with many levels, without obfuscating meaning. These tree-based methods have been used successfully in many clinical settings.

Prediction performance and clinical insight are thus sometimes complementary but more often at odds with each other. These 2 concepts should not be confused, and when reporting results the original goal of the analysis should dictate not only the appropriate method but also the format for reporting the results to avoid misunderstanding.