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 nonevents) 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 colleagues[2] 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 colleagues[2] 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 classing 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 colleagues[2] 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.

The authors reported no conflicts of interest.

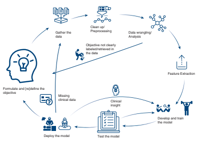
The *Journal* policy requires editors and reviewers to disclose conflicts of interest and to decline handling or reviewing manuscripts for which they may have a conflict of interest. The editors and reviewers of this article have no conflicts of interest.

Hemant Ishwaran, PhD[a]
Robert O'Brien, PhD[b,c]
[a]*Division of Biostatistics*
University of Miami
Miami, Fla
Departments of [b]*Data Science*
[c]*Surgery*
University of Mississippi Medical Center
Jackson, Miss

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.e8.
2. Nedadur R, Tam DY, Fremes SE. Machine learning and readmission: do we need new methods to solve old problems? *J Thorac Cardiovasc Surg*. 2022;163:e101-2.
3. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*. 2015;10:e0118432.
4. Flach P, Kull M. Precision-recall-gain curves: PR analysis done right. In: Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R, eds. *NIPS 15: Proceedings of Advances in Neural Information Processing Systems 28*. Cambridge (MA): MIT Press; 2015:838-46.
5. Halligan S, Altman DG, Mallett S. Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: a discussion and proposal for an alternative approach. *Eur Radiol*. 2015;25:932-9.
6. Spiegelhalter DJ. Probabilistic prediction in patient management and clinical trials. *Stat Med*. 1986;5:421-33.
7. Breiman L. Random forests. *Machine Learn*. 2001;45:5-32.
8. Ishwaran H, Lu M. Standard errors and confidence intervals for variable importance in random forest regression, classification, and survival. *Stat Med*. 2019;38:558-82.
9. Friedman J. Greedy function approximation: a gradient boosting machine. *Ann Stat*. 2001;29:1189-232.

https://doi.org/10.1016/j.jtcvs.2020.07.113



REPLY: IN MACHINE LEARNING, THE DEVIL IS IN THE DETAILS

Reply to the Editor:

We thank Nedadur and colleagues[1] for their interest and insightful comments in our study of using machine learning in thoracic surgery outcomes, specifically predicting early readmission after esophagectomy. We agree with the message of the letter that comprehensive reporting of machine-learning models is key to understanding the inner working of the model and the conclusions reached. However, as it was pointed out by Ishwaran and O'Brien,[2] we diverge from their emphasis on performance measures rather than methodology and the clinical applications. Given the variable nature of prediction models and the uneven importance of metrics for each clinical question, while many guidelines have been developed for reporting performance metrics, there is inconsistency and debate on what is important to report. We further caution against using any published machine-learning methods/schema to actual clinical practice without externally validating them. As we learned after further investigation of our work, even if all the steps of data-oriented research including problem formulation, design/development of methodology, data gathering, and preparation are solid, tree-based models are very sensitive in their experimental evaluation and a minor error may affect the obtained results significantly. Thankfully, while the error overestimated the area under the curves and other metrics of the models, it did not affect the conclusion of our work. The random forest model was the most accurate model among those presented; after applying NearMiss method, the sensitivity of the model increases, which comes at the cost of decreasing specificity and accuracy. Both models have superior predictive ability than logistic regression, and order of blended metrics presented in Table E7 of our article remained intact.

The issue was raised that if the calibration metrics are not reported, then the probability of certainty is not clear. Our main focus in this work was discrimination, and we did not focus on calibration of the model. As the authors pointed out, calibration plots will show the reliability of predictions as a function of change in the prevalence. For the sake of completeness, we take this opportunity to present the calibration plots for the classifier used along with isotonic and sigmoid calibration[3] of the model (Figure 1) and the corresponding Brier scores,[4] precisions, recalls, and F1 scores[5] (Table 1).

This discussion gives us an opportunity to present a measure that is sometimes overlooked in prediction models of biomedical sciences: the precision–recall curve. In our case, the curves capture how much precision is traded between the models to gain the recall (sensitivity) (Figure 2). As we argued in work, we believe in the case of clinical decisions, this is often necessary (in our case, we would like to catch as many patients that are to be readmitted early after discharge, even at the cost of losing much of the precision). The best way to measure this among models is by a weighted blended metric like F_β score, which is the weighted harmonic mean of precision and recall. This measure can be used to compare the 2 models when precision and recall (sensitivity) have different importance. That is, if we value recall (sensitivity), β times as much the precision. F_β score is often used to compare models when false negatives have greater detriment than false positives