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}

^aDivision of Biostatistics

University of Miami

Miami, Fla

Departments of ^bData Science

^cSurgery

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 colleagues¹ 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 calibration³ of the model (Figure 1) and the corresponding Brier scores,⁴ precisions, recalls, and F1 scores⁵ (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

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.

and false positives can be tolerated (which as we argued, is the case for catching early readmissions after esophagectomy). In our measurements, we assigned $\beta = 4$, which was an arbitrary choice; however, we think in the clinical environment, β should be assigned a greater value to be able to capture the detriment of false negatives compared to false positives.

Another issue was that to appropriately compare the performance of machine learning and regression, we must be able to compare the same performance metrics. Given that we provided the confusion matrix for all of our models, the metrics that are not mentioned in the body of the text can be estimated for every model (for example,

specificity = true negatives / (true negatives + false positives), the mean of all of which can be observed in the confusion matrix). However, this is not an exact mean of the measure since, if metrics are calculated using the mean of the component of the confusion matrix, it will not be the same as the mean of that metric over all folds. The reason that we chose to present accuracy in the body of the text for logistic regression is because it is notorious to have a bias toward majority class,^{6,7} which makes specificity a fraught measure.

Furthermore, in their letter the authors asked how we handle class imbalance. We discuss this issue briefly in the body of the article and in some detail in the Online Data Supplement. In Table E1 of the article, we explain our rationale for using NearMiss subsampling to deal with class imbalance. NearMiss is a well-known undersampling method used to deal with imbalanced data. This undersampling method selects majority class instances with minimum distance to each minority class instance.⁸ Furthermore, there is a commentary of our work by Ishwaran and O'Brien⁹ about class imbalances, how we dealt with it in our work, and what other methods are there that

FIGURE 1. Calibration plots (reliability curve) of the models of early readmission of esophagectomy patients. The calibration is based on the precision probability (using *predict_proba* of python). For creating the plots, *sklearn.calibration.CalibratedClassifierCV* of python was used. Sigmoid calibration method corresponds to Platt's calibration method.³ Isotonic regression corresponds to nonparametric isotonic regression calibration. Brier score corresponding to each reliability curve is shown in the figure legend in *parentheses*.

TABLE 1. Brier score, precision, recall, and F1 score for reliability plots shown in Figure 1

Metric	Uncalibrated RandomForest	Isotonic calibrated RandomForest	Sigmoid calibrated RandomForest	Logistic regression
Brier score	0.118	0.115	0.114	0.135
Precision	0.808	0.625	0.614	0.667
Recall	0.216	0.206	0.278	0.082
F1 score	0.341	0.310	0.383	0.147

can ameliorate imbalance problems in the datasets for interested readers.

In summary, we recognize that when it comes to machine learning, most of the work is in the details. While we made all the necessary measures and metrics for our models available to the readers, given the nature of the *Journal* having mostly clinical readership, we chose to provide some of the details in the Online Data Supplement. Furthermore, we acknowledge that a prediction model is only as good as data it uses to train and predict. The database we used for building our prediction model is a national database, not a granular one. This lack of granularity affects the model greatly, and we agree with Altorki and Sedrakyan in their commentary that these limitations make the model not ready for “prime time.”¹⁰

However, the methodology provided in the article can be used and expanded on a more granular database to achieve a model ready to be used in the “front end.” Finally, we agree that machine learning cannot be used as a scoring system at the bedside. In the article, we provided a schema so that the electronic health records can be used to alert the team at the time of discharge that a patient is high risk for early readmission.¹¹ We also share the sentiment that until the investment is made in prospective studies using surgeon-friendly tools to adopt machine learning frameworks, the potential of these systems will not be fully actualized.

Siavash Bolourani, MD^{a,b,c,d}
Theodoros P. Zanos, PhD^d
Ping Wang, MD^{a,b}

FIGURE 2. PR curves of models of early readmission of patients undergoing esophagectomy. PR curves for our random forest predictive model before (A) and after (B) application of the NearMiss algorithm and compared with logistic regression (C). To show low bias, we used 4-fold cross validation by splitting the dataset into 4 consecutive folds and all permutations choosing 3 of the folds out of 4 were used as a training set (ie, 4 distinct permutations of folds). StratifiedKFold of python was used to perform validation and depict the resulting PR curves. Mean PR curve and ± 1 standard deviation depicted in blue and shaded areas around it, respectively. PR, Precision–recall curve; AUCPR, area under the precision–recall curve.

Mohammad A. Tayebi, PhD^e

Paul C. Lee, MD, MPH^c

^aElmezzi Graduate School of Molecular Medicine

^bDepartment of Surgery

^cDepartment of Cardiovascular and Thoracic Surgery,

Donald and Barbara Zucker School of Medicine at Hofstra/

Northwell

^dInstitute of Bioelectronic Medicine

Feinstein Institutes for Medical Research

Northwell Health

Manhasset, NY

^eSchool of Computing Science

Simon Fraser University

Burnaby, British Columbia, Canada

References

1. 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.
2. Ishwaran H, O'Brien R. Reply: the standardization and automation of machine learning for biomedical data. *J Thorac Cardiovasc Surg.* 2022;163:e102-3.
3. Lin H-T, Lin C-J, Weng RC. A note on Platt's probabilistic outputs for support vector machines. *Machine Learn.* 2007;68:267-76.
4. Benedetti R. Scoring rules for forecast verification. *Monthly Weather Rev.* 2010;138:203-11.
5. Goutte C, Gaussier E. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. *Lecture Notes Computer Sci.* 2005;3408:345-59.
6. Muchlinski D, Siroky D, He J, Kocher M. Comparing random forest with logistic regression for predicting class-imbalanced Civil War onset data. *Polit Anal.* 2016;24:87-103.
7. Abd Elrahman SM, Abraham A. A review of class imbalance problem. *J Netw Innovative Comput.* 2013;1:332-40.
8. Lemaitre G, Nogueira F, Aridas CK. Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. *J Mach Learn Res.* 2017;18:1-5.
9. Ishwaran H, O'Brien R. Commentary: the problem of class imbalance in biomedical data. *J Thorac Cardiovasc Surg.* 2021;161:1940-1.
10. Altorki N, Sedrakyan A. Commentary: Can machine learning reduce readmissions after esophagectomy? A consummation devoutly to be wished. *J Thorac Cardiovasc Surg.* 2021;161:1944-5.
11. 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.

<https://doi.org/10.1016/j.jtcvs.2020.10.052>