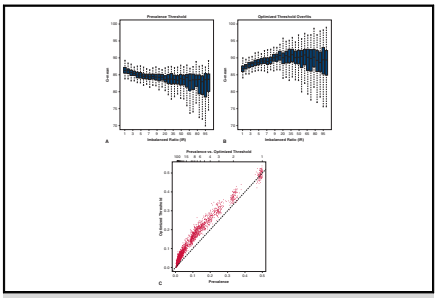
See Article page 1433.

Commentary: To classify means to choose a threshold

Jiangnan Lyu, BSc, and Hemant Ishwaran, PhD



The prevalence threshold yields accurate classification without dangerous data snooping.

CENTRAL MESSAGE

Classification requires a threshold; however, methods like C-statistic and AUC obfuscate this. Luckily, there is a sensible strategy for imbalanced data thresholding.

Movahedi and colleagues[1] point out that precision recall area under the curve (PR-AUC) can be a better performance-evaluation tool than receiver operating characteristic area under the curve (ROC-AUC) for imbalanced data. This same point has also been made in recent editorials.[2,3] By comparing ROC and PR applied in a 90-day left ventricular assist device mortality study, the authors[1] conclude that ROC fails to reflect a classifier's performance in detecting the rare cases by generating overly optimistic AUC. While we generally agree with this message, we wish to clarify certain points concerning classification and to note some recent developments.

Soft classification[4] is the problem of classifying an object using probability. The ubiquitous Bayes classifier assigns an object to 1 of 2 groups if probability exceeds 0.5. For machine learning (ML) methods, this often results in nearly all cases being classified to the majority group when data are highly imbalanced[5] (in the authors' study, 92% of patients survive, the majority group, 8% die, the minority group; a relatively high imbalanced ratio [IR] of 92/8 = 11.5). The value 0.5 used by the Bayes classifier is called the threshold, and without such a threshold, soft classification cannot be performed.

ROC-AUC is insensitive to IR. Such a property is unwanted for imbalanced data, since rare cases are usually associated with greater costs; proper performance metrics should show a monotonic decrease with increasing IR. While PR-AUC has this property, making it more suitable for imbalanced data, both methods fail to address soft classification. AUC methods like these provide an overall measure of performance by varying a hypothetical threshold but are silent on actual threshold value needed for soft classification.

There is a simple solution, called q*-classification, designed specifically for imbalanced data.[5,6] This replaces the 0.5 threshold used by the Bayes classifier with the prevalence (fraction of minority group to overall sample size). Figure 1 shows G-mean (geometric mean; an appropriate metric for imbalanced data) soft classification performance for the ML method random forest (RF). In Figure 1, *A*, RF uses q*-classification thresholding: performance is excellent, even with extreme imbalanced data, IR = 100. In Figure 1, *B*, RF uses threshold maximizing cross-validated G-mean: while performance appears excellent, results are optimistically biased due to overtraining data (notice G-mean improves with worsening IR). Figure 1, *C*, shows optimized threshold is inflated compared with prevalence values. Taken together, this shows superiority of the prevalence threshold without dangers of overtraining.

In conclusion, the authors work adds to the growing concern of the misuse of ROC and C-statistics with imbalanced data. To their credit, the authors identify soft classification and the issue of threshold selection as a limitation of their study and call for future studies to address this. However, we caution that informal strategies to select threshold values may be doomed by the dangers of data snooping, which is exacerbated by the challenges of imbalanced data. We recommend q*-classification, which is an easily

From the Division of Biostatistics, Miller School of Medicine, University of Miami, Miami, Fla.

Disclosures: 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.

Received for publication July 30, 2021; revisions received July 30, 2021; accepted for publication Aug 3, 2021; available ahead of print Aug 8, 2021.

Address for reprints: Hemant Ishwaran, PhD, Don Soffer Clinical Research Center, University of Miami, 1120 NW 14th St, Miami, FL 33136 (E-mail: hishwaran@med.miami.edu).

J Thorac Cardiovasc Surg 2023;165:1443

0022-5223/$36.00

https://doi.org/10.1016/j.jtcvs.2021.08.009

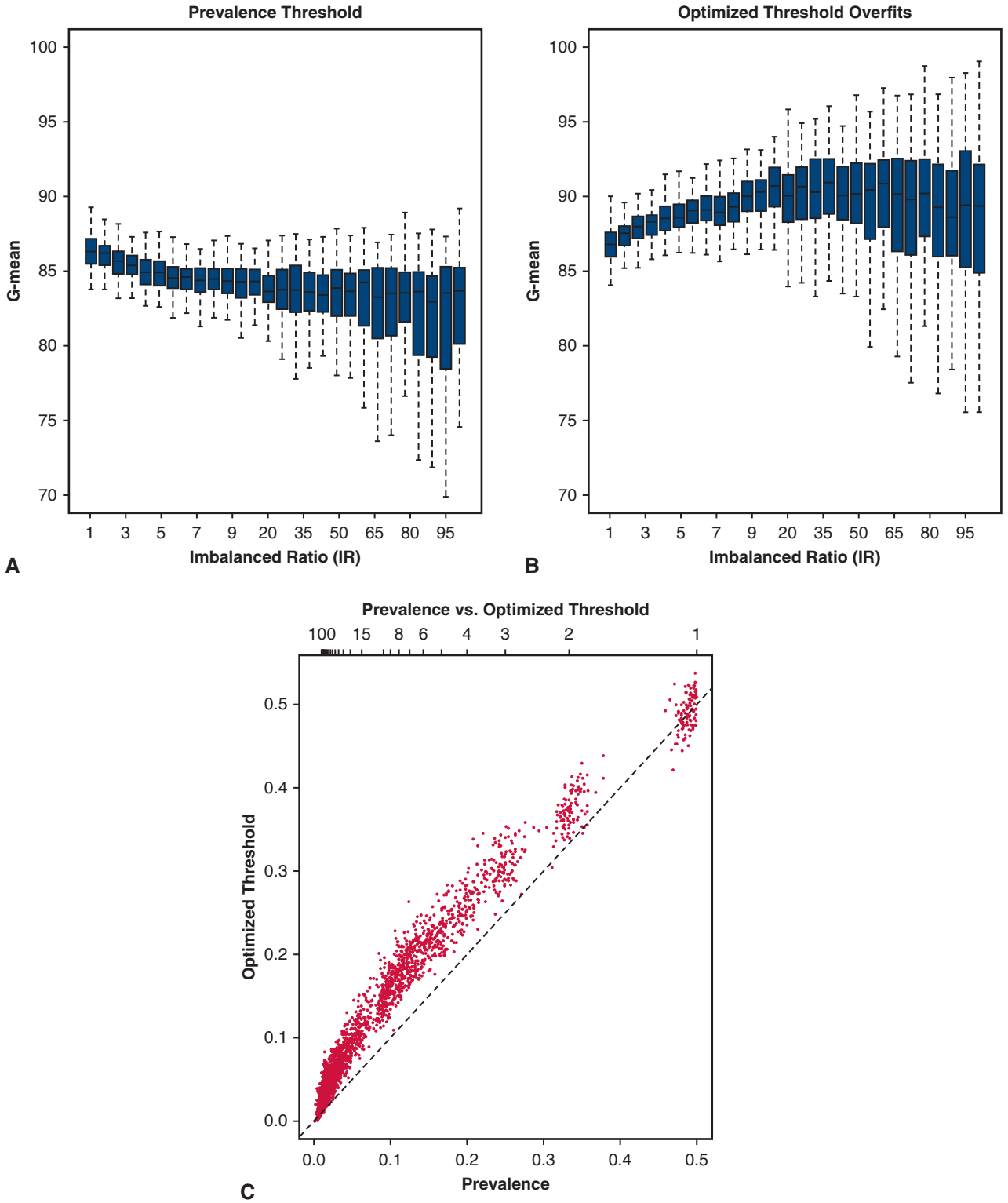


FIGURE 1. G-mean (geometric mean) soft classification performance of the machine learning method random forest (RF). Data are classified as a rare case if RF out-of-bag (cross-validated) probability is larger than a specific threshold value. Classification data were simulated 100 times independently under imbalanced ratio (IR) varying from balanced (IR = 1) to extreme imbalanced (IR = 100) scenarios. A, Threshold for RF classification equals prevalence (fraction of rare cases), a method called RFQ.[6] Performance of RFQ is excellent across all IR values. B, Threshold for RF classification is selected by maximizing out-of-bag (cross-validated) G-mean. Even though optimization uses cross-validated values, results are optimistically biased as evident by G-mean values increasing with IR. C, Optimized threshold values are inflated when compared with prevalence threshold values (the only exception being IR = 1 when data are balanced; *top right*). Combined, this demonstrates optimality of RFQ (q*-classification) while avoiding double-dipping the data.

calculated threshold value, with guaranteed theoretical properties.[6] When combined with a flexible ML method like RF, this yields excellent performance.

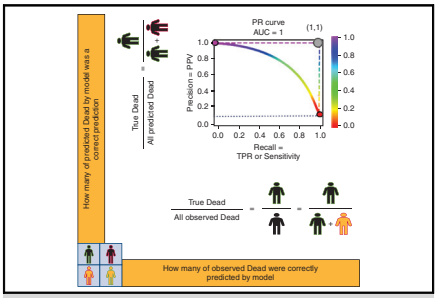
References

1. Movahedi F, Padman R, Antaki JF. Limitations of receiver operating characteristic curve on imbalanced data: assist device mortality risk scores. *J Thorac Cardiovasc Surg*. 2023;165:1433-42.e2.
2. Ishwaran H, O'Brien R. Reply: the standardization and automation of machine learning for biomedical data. *J Thorac Cardiovasc Surg*. 2021;165:1433-42.
3. Ishwaran H, Blackstone EH. Commentary: dabblers: beware of hidden dangers in machine-learning comparisons. *J Thorac Cardiovasc Surg*. 2021;165:1433-42.
4. Wahba G. Soft and hard classification by reproducing kernel Hilbert space methods. *Proc Natl Acad Sci USA*. 2002;99:16524-30.
5. Ishwaran H, O'Brien R. Commentary: the problem of class imbalance in biomedical data. *J Thorac Cardiovasc Surg*. 2021;161:1940.
6. O'Brien R, Ishwaran H. A random forests quantile classifier for class imbalanced data. *Pattern Recogn*. 2019;90:232-49.

See Article page 1433.

Commentary: Machine learning and the brave new world of risk model assessment

Paul Kurlansky, MD



Precision recall curve.

CENTRAL MESSAGE

Machine learning challenges traditional methods of risk model assessment. Emphasizing the less-common outcome in an unbalanced dataset may be more appropriate than conventional approaches.

Cardiac surgeons have become leaders in the development and implementation of well-constituted risk models.[1] The relatively concrete nature of our profession—discrete events and outcomes—facilitates quantifiable risk modeling, robust risk adjustment, and meaningful attribution. Traditional approaches have relied on increasingly sophisticated statistical modeling techniques.[2] The introduction of machine learning with its remarkable ability to identify frequently unrecognized associations and patterns in large data sets will have an increasingly profound influence on the development and application of predictive modeling.[3] The various approaches—supervised versus unsupervised, classification/regression versus clustering, and their various subsets—have particular strengths, weaknesses, and optimal applications. Although a welcome and refreshing development, this rapidly evolving field requires careful attention to the fundamental question: Are these approaches actually better than what we already have? Appealing is not necessarily better. How do we know? It is specifically in answer to this question that Movahedi and colleagues[4] provide a very patient and thorough tutorial examining the challenge of optimal model assessment in the face of unbalanced data. Specific context is provided by the comparison of the logistic regression-based HeartMate Risk Score[5]—derived and validated from the clinical trial data of 1122 patients with a left ventricular assist device—with a random forest plot-based machine learning approach derived from large multicenter registry data. The point is not which model performed better, but rather how to best make that determination. Virtually all cardiac surgical risk models evaluate relatively infrequent events. Therefore, there are many more true negatives than true positives—an unbalanced sample. The classic

From the Division of Cardiac Surgery, Department of Surgery, Columbia University, New York, NY.

Disclosures: The author 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.

Received for publication Aug 4, 2021; revisions received Aug 4, 2021; accepted for publication Aug 9, 2021; available ahead of print Aug 14, 2021.

Address for reprints: Paul Kurlansky, MD, Division of Cardiac Surgery, Department of Surgery, Columbia University, Neurological Institute 554, 710 W 168th St, New York, NY 10032 (E-mail: pk2245@cumc.columbia.edu).

J Thorac Cardiovasc Surg 2023;165:1445-6
0022-5223/$36.00

https://doi.org/10.1016/j.jtcvs.2021.08.029