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

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

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.

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 colleagues⁴ 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 Score⁵—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

Copyright © 2021 by The American Association for Thoracic Surgery

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

FIGURE 1. Precision recall curve. PPV, Positive predictive value; TPR, true positive rate.

tool for evaluation, the receiver operator characteristics—a balance that plots sensitivity versus 1-specificity—is intuitively appealing. Both sensitivity and specificity are characteristics of the model, independent of the prevalence in the population and therefore might seem ideal for model assessment. However, what we really care about is whether or not the model says that a certain outcome; for example, mortality is probable, will it really happen? (Similar to a diagnostic test—if the test is positive, does the patient really have the disease) To answer that question—precision, or positive predictive value—one must account for the

prevalence of disease in the population: the higher the prevalence, the greater the precision. True positives, which assume a predominant role in both sensitivity (true positive/true positive + false negative) and precision (true positive/true positive + false positive) drop completely from the equation in specificity (true negative/true negative + false positive). Therefore, in an unbalanced population that has relatively few true positives, a technique that highlights those data by accounting for sensitivity and precision—the precision recall (sensitivity) curve (see Figure 1)—can more readily identify the value of a risk model in a population with the low prevalence of the outcome of interest, as is quite graphically demonstrated in this provocative tutorial. Obviously, no single test is adequate, and any test of discrimination must be balanced with careful assessment of calibration; that is, how closely do the predicted outcomes correspond to those observed? However, with new approaches to risk assessment come new approaches to model assessment.

References

1. Shahian DM, Fernandez FG, Badhwar V. The Society of Thoracic Surgeons National database at 30: honoring our heritage, celebrating the present, evolving for the future. *Ann Thorac Surg.* 2019;107:1259-66.
2. O'Brien SM, Feng L, He X, Xian Y, Jacobs JP, Badhwar V, et al. The Society of Thoracic Surgeons 2018 adult cardiac surgery risk models: part 2-statistical methods and results. *Ann Thorac Surg.* 2018;105:1419-28.
3. Blackstone EH, Swain J, McCardle K, Adams DH, Governance Committee, American Association for Thoracic Surgery Quality Assessment Program. A comprehensive American Association for Thoracic Surgery Quality Program for the 21st century. *J Thorac Cardiovasc Surg.* 2019;158:1120-6.
4. 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.
5. Cowger J, Sundareswaran K, Rogers JG, Park SJ, Pagani FD, Bhat G, et al. Predicting survival in patients receiving continuous flow left ventricular assist devices: the HeartMate II risk score. *J Am Coll Cardiol.* 2013;61:313-21.