See Article page 2075.

ADULT

Commentary: Dabblers: Beware of hidden dangers in machine-learning comparisons

Hemant Ishwaran, PhD,[a] and
Eugene H. Blackstone, MD[b]

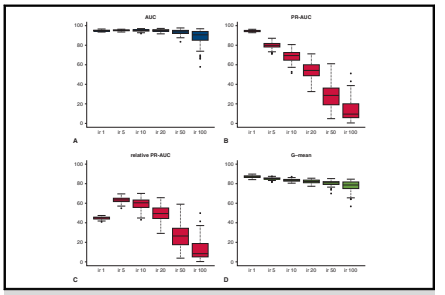
Readers of Benedetto and colleagues' article[1] promising that machine learning (ML) improves risk prediction after cardiac surgery should be underwhelmed.[1] No ML method was found better at discrimination than logistic regression; only by selecting the best-performing method from each article was there a single $P = .03$. Disappointing. This finding was blamed on date of publication: From 1997 to 2009, only 2 of 6 papers showed improvement by any ML method over logistic regression, versus 6 of 9 from 2010 to 2018. Other factors play a role.

SINGLE INSTITUTION

Twelve of the 15 articles are single-institution studies. They had relatively small numbers of patients and events, possibly leading to poor ML and logistic regression results. The 3 large database studies (published before 2010[2-4]) showed no improvement.

RESULTS OF CARDIAC SURGERY WERE IMPROVING

With improved results, there are fewer events. Consequently, data become more imbalanced (nonevents outweigh events). Imbalance was not considered in the 1990s and 2000s, and those dabbling with ML in a nonrigorous fashion today do not recognize this either. This is a huge challenge for ML and unfortunately has led to numerous ad hoc methods for dealing with it. Many of these have no theoretical justification, and naïve users may not understand their implications. One popular oversampling technique manufactures new data to balance the frequency of events; for highly imbalanced data, this involves manufacturing a large amount of data and may lead to false results.



Misleading C-statistics (AUC) versus precision recall (PR-AUC) and G (geometric) mean for imbalanced data.

CENTRAL MESSAGE

Machine learning is not for dabblers. An underappreciated problem is imbalance between number of events and nonevents, for which traditional *C*-statistics are an inappropriate evaluation metric.

THE WRONG METRIC

The metric used for assessing improvement of ML over logistic regression was the *C*-statistic (area under curve). In the 15 studies considered, all reported relatively few events (range, 3%-25.5%; median, 6.3%; mean, 8.1%), an extreme class imbalance. The *C*-statistic is inappropriate because it neither accounts for frequency of events nor unequal misclassification costs. More appropriate metrics are precision-recall curves and geometric mean indices. High precision means a high positive predictive value, and high recall a high sensitivity; geometric mean measures the balance of true positive and true negative rates (Figure 1).[5]

CLINICAL SIGNIFICANCE

To be meaningful clinically, calibrated probabilities of events are essential. Many ML methods cannot deliver these. Instead, they provide a continuous value for discrimination. This makes it possible to evaluate classification

From the [a]Division of Biostatistics, University of Miami, Miami, Fla; and [b]Department of Thoracic and Cardiovascular Surgery, Heart, Vascular, and Thoracic Institute; and Department of Quantitative Health Sciences, Lerner Research Institute, Cleveland Clinic, Cleveland, Ohio.

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 Aug 24, 2020; revisions received Aug 24, 2020; accepted for publication Aug 26, 2020; available ahead of print Aug 31, 2020.

Address for reprints: Eugene H. Blackstone, MD, Department of Thoracic and Cardiovascular Surgery, Heart, Vascular, and Thoracic Institute, Cleveland Clinic, 9500 Euclid Ave, JJ-4, Cleveland, OH 44195 (E-mail: blackse@ccf.org).

J Thorac Cardiovasc Surg 2022;163:2088-90
0022-5223/$36.00

https://doi.org/10.1016/j.jtcvs.2020.08.091

performance (which is why *C*-statistics are used) but not actual patient risk. Even for ML methods providing probabilities, *C*-statistics can be high but calibration poor in highly imbalanced data. Also, actual threshold value for classifying patients is not provided by many ML methods, which makes them impractical for clinical use.

LIMITED VARIABLES

Cardiac surgery is more than 60 years old and during that time has established risk factors for postoperative adverse outcomes. However, nearly all models for these outcomes are parsimonious, and unusual factors or combinations of factors do not budge *C*-statistics, so are eliminated. Interactions among factors have received scant attention. But that

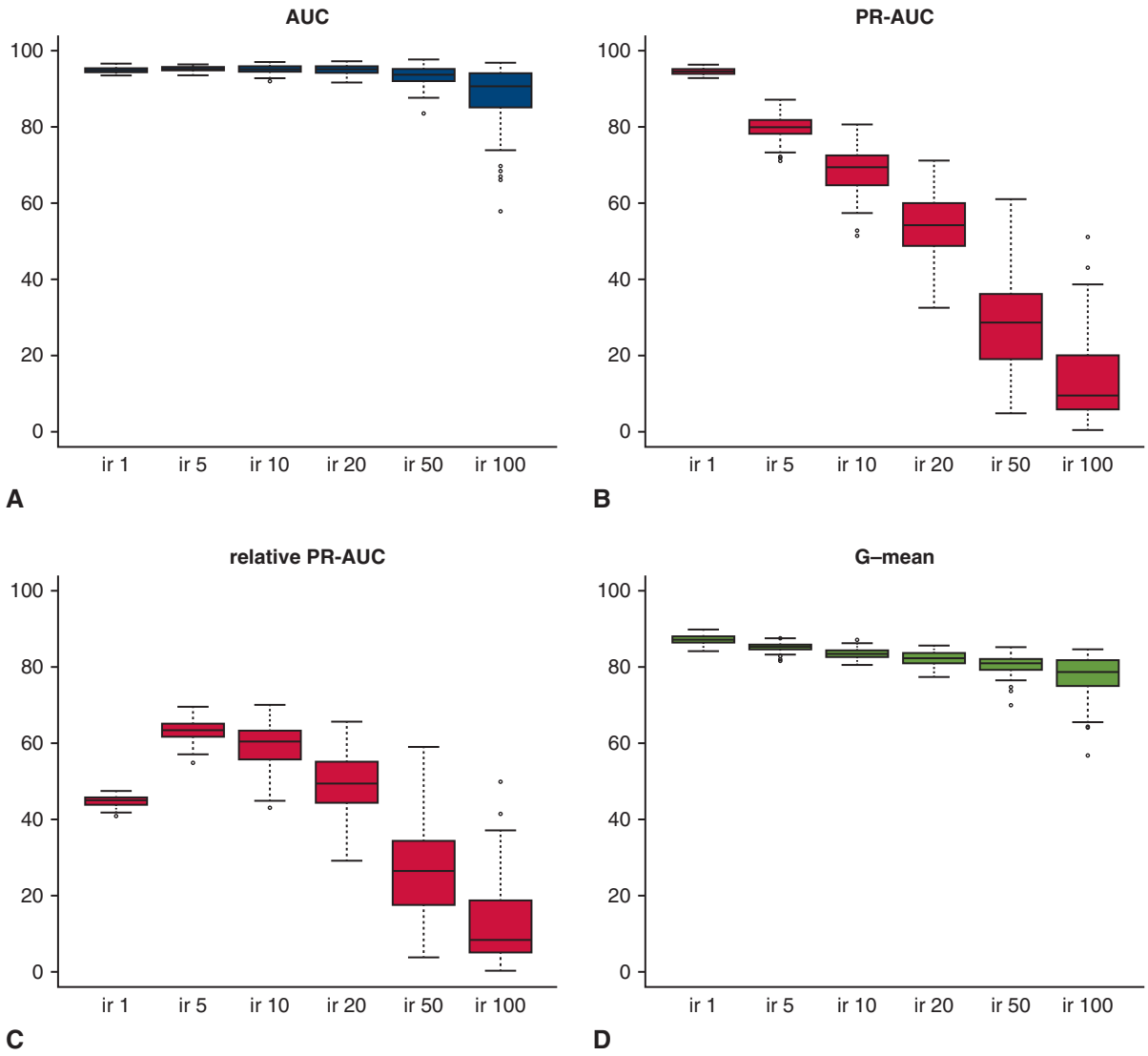


FIGURE 1. Performance of random forest quantile classifier (RFQ), a machine-learning method that makes few assumptions, using appropriate metrics for imbalanced data (few events compared with nonevents, as is typical of most complications after cardiac surgery), contrasted with inappropriate *C*-statistics (area under curve [AUC]) in such settings. Classification data were simulated 100 times independently under imbalanced ratios (IR; *ir* in figures) of 1 (IR = 1 along horizontal axis, balanced number of nonevents to events) and IRs of 5, 10, 20, 50, and 100—moderate, high, to extreme imbalance, corresponding, for example, to mortality of 16.7% (IR = 5), 9.1% (IR = 10), 4.8% (IR = 20), 2.0% (IR = 50), and 1.0% (IR = 100). *Box* encompasses the 25th and 75th percentiles of values; *thick horizontal line* is the median value, and *whiskers* extend 1.5 times the interquartile range beyond where the box ends. A, *C*-statistic, AUC of receiver-operator curve (ROC) (sensitivity vs 1 – specificity), is nearly constant across IRs, dipping down only slightly for extreme ratio of IR = 100, leading to false assessment of classifier performance. B, Precision recall AUC (PR-AUC) denotes precision—positive predictive value—vs recall—sensitivity; values decrease with increasing imbalance. C, PR-AUC must be calibrated to IR; here PR-AUC is subtracted by PR-AUC for random classifier, resulting in relative PR-AUC. It shows correct decrease in performance with increasing imbalance from IR = 5 to IR = 10, 20, 50, and 100. D, G-mean performance shows correct decrease in performance with increasing IR.

is not where ML shines. Real ML comes into its own when exposed to a large number and type of variables, including those that have not been considered, or new ones like omics on micro and macro scales.

ML IS DEVELOPING FAST

Although ML's history matches that of cardiac surgery, its methodology development is inverse: Cardiac surgery was explosive in its early years; ML is only now experiencing explosive growth.[6] Surgeons seriously interested in ML need to team up with professionals developing a field that has moved past amateur dabblers.

References

1. Benedetto U, Dimagli A, Sinha S, Cocomello L, Gibbison B, Caputo M, et al. Machine learning improves mortality risk prediction after cardiac surgery: systematic review and meta-analysis. *J Thorac Cardiovasc Surg*. 2022;163:2075-87.e9.
2. Tu JV, Weinstein MC, NcNeil BJ, Naylor D, the Steering Committee of the Cardiac Care Network of Ontario. Predicting mortality after coronary artery bypass surgery: what do artificial neural networks learn? *Med Decis Making*. 1998;18:229-35.
3. Lippmann RP, Shahian DM. Coronary artery bypass risk prediction using neural networks. *Ann Thorac Surg*. 1997;63:1635-43.
4. Nilsson J, Ohisson M, Thulin L, Hoglund P, Nashef SA, Brandt J. Risk factor identification and mortality prediction in cardiac surgery using artificial neural networks. *J Thorac Cardiovasc Surg*. 2006;132:12-9.
5. Akosa JS. Predictive accuracy: a misleading performance measure for highly imbalanced data. Accessed August 19, 2020. Available at: https://support.sas.com/resources/papers/proceedings17/0942-2017.pdf
6. Efron B, Hastie T. *Computer Age Statistical Inference*. Cambridge, United Kingdom: Cambridge University Press; 2016.

See Article page 2075.

Commentary: Machine learning and cardiac surgery risk prediction

David M. Shahian, MD,[a] and
Richard P. Lippmann, PhD[b]

David M. Shahian, MD (*left*), and Richard P. Lippmann, PhD (*right*)

CENTRAL MESSAGE

Machine learning is only modestly superior to logistic regression for prediction of cardiac surgery mortality, possibly because of low-dimensional predictors with weak nonlinear relationships.

In 1997, we published the first study comparing coronary artery bypass grafting mortality risk prediction using standard logistic regression versus what was then a state-of-the-art multilayer perceptron neural network, a type of machine learning.[1] The c-indices (receiver operating characteristic curve areas) were nearly identical (0.76) for logistic regression, neural networks, and a committee or ensemble classifier that combined estimates from the other 2 approaches, although the committee classifier had slightly better calibration. We hypothesized that these findings might indicate "absence of complex nonlinear relationships, at least among the variables presented to the network," the latter caveat emphasizing the limitations in the available predictor variables.

What has happened during the 23 years since our original study? Given the availability of newer machine learning approaches and vast improvements in computer memory and processing speeds, are there now more convincing demonstrations of the superiority of machine learning for cardiac surgery risk prediction?

From the [a]Division of Cardiac Surgery, Department of Surgery, and Center for Quality and Safety, Massachusetts General Hospital, Boston; and [b]Massachusetts Institute of Technology Lincoln Laboratory, Lexington, Mass.

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 Aug 18, 2020; revisions received Aug 18, 2020; accepted for publication Aug 18, 2020; available ahead of print Aug 24, 2020.

Address for reprints: David M. Shahian, MD, Division of Cardiac Surgery, Department of Surgery, and Center for Quality and Safety, Massachusetts General Hospital, 55 Fruit St, Boston, MA 02114 (E-mail: dshahian@partners.org).

J Thorac Cardiovasc Surg 2022;163:2090-2
0022-5223/$36.00

https://doi.org/10.1016/j.jtcvs.2020.08.058