http://dx.doi.org/10.2106/JBJS.21.00341

January 6, 2022

Which model can help outcome prediction in orthopedic trauma surgery remains a problem

Ling-xiao He

Department of Orthopedics, West China Hospital, Sichuan University, Chengdu, China

Other Contributors:

Ning Ning

Department of Orthopedics, West China Hospital, Sichuan University, Chengdu, China

Zong-ke Zhou

Department of Orthopedics, West China Hospital, Sichuan University, Chengdu, China

We have carefully read the paper by Jacobien et al. published in JBJS comparing machine learning(ML) methods with logistic regression(LR) methods. The study compared the effectiveness of LR with penalized logistic regression, support vector machine, decision tree algorithms, neural network, and Bayesian classifiers in datasets of different sample sizes and disease types, finding that the difference in performance between ML models was insignificant and not superior to LR models.

We agree that ML methods may not be better than traditional regression analysis in medical data, and their performance relies on large datasets. However, we would like to discuss some problems in this study that may affect model construction and model performance.

First, the authors performed variable selection before modeling with variable selection methods, including LASSO, Boruta, and random forest recursive selection algorithms. Did all of these methods select the same subsets of variables? Besides, if there are significant differences between these methods, what criteria did the authors adopt to rank the importance of the selected variables? The above two points are critical because variable selection and model training will mutually determine the final model performance (1). Different variable selection methods have different mathematic rationales, and the model performance accuracy may not be optimal when combining a model with an inappropriate feature selection method.

Second, there is a certain degree of sample imbalance in each dataset, with the incidence of outcomes

ranging from 4.5% to 44%. While the AUC, or C statistic, is usually suitable for assessing the discrimination performance of a model on balanced samples (e.g., the binary data with comparable positive and negative sample sizes), but it does not consider either the frequency of events or the unequal misclassification costs (2). So, we suggest choosing other metrics to evaluate discrimination focusing on the proportion of positive events, such as the F1 score, or comparing Precision-Recall curves of different models (3).

Third, the authors did not specify the parameters used in different models that may affect the accuracy of the performance. When using machine learning, the parameters need to be adjusted appropriately according to the specific dataset, and the optimal model should be output based on repeated model training. We suggest that researches related to machine learning should be conducted under the premise of disclosing the algorithms and critical parameters. Otherwise, it will not be conducive to independent external validation and further revision of the model (4).

Fourth, the paper compared model performance metrics but did not show the “best model” criteria. And the reasons for selecting these models are also not listed in the article. Van Calster et al. (5) suggested that poorer calibrations will have greater impacts on the clinical utility of prediction models than poorer AUC metrics. Demonstrating the metrics and performance parameters for screening the best models in the training sets may help understand this research.

In conclusion, we agree that LR combining empirical and statistical properties for medical data is comparable to the results of current emerging ML methods in many cases. Still, ML methods have their methodological uniqueness and potential for further optimization. In the future, the methods and parameters should be determined more carefully in model construction to increase the possibilities of more accurate predictions.

Disclaimer: e-Letters represent the opinions of the individual authors and are not copy-edited or verified by JBJS.

References

1. Remeseiro B, Bolon-Canedo V. A review of feature selection methods in medical applications. Comput Biol Med. 2019 Sep;112:103375.
2. Ishwaran H, Blackstone EH. Commentary: Dabblers: Beware of hidden dangers in machine-learning comparisons. J Thorac Cardiovasc Surg. 2020 Aug 31 [Epub ahead of print].
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 Mar 4;10(3):e0118432.
4. Van Calster B, Wynants L, Timmerman D, Steyerberg EW, Collins GS. Predictive analytics in health care: how can we know it works? J

Am Med Inform Assoc. 2019 Dec 1;26(12):1651-1654.
5. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW; Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019 Dec 16;17(1):230.

Conflict of Interest: None Declared