ESC European Society of Cardiology
European Heart Journal - Digital Health (2024) **5**, 1–2
https://doi.org/10.1093/ehjdh/ztad059

LETTER TO THE EDITOR

Challenges in developing and validating machine learning models for transcatheter aortic valve implantation mortality risk prediction

Sina Kazemian[1,2], Mahbod Issaiy[3], and Kaveh Hosseini [1,2,*]

[1]Cardiac Primary Prevention Research Center, Cardiovascular Diseases Research Institute, Tehran University of Medical Sciences, Tehran Heart Center, Kargar St. Jalal al-Ahmad Cross, 1411713138, Tehran, Iran; [2]Tehran Heart Center, Cardiovascular Diseases Research Institute, Tehran University of Medical Sciences, Imam Khomeini Hospital Complex, Tohid Square, 1419733141, Tehran, Iran; and [3]Advanced Diagnostic and Interventional Radiology Research Center (ADHR), Tehran University of Medical Sciences, Tehran, Iran

Online publish-ahead-of-print 11 October 2023

Commentary article to: 'Challenges in developing and validating machine learning models for TAVI mortality risk prediction: reply', by A. Leha *et al.*, https://doi.org/10.1093/ehjdh/ztad065.

We read with interest the article by Leha *et al.*,[1] developing and validating the TRIM risk scores for predicting the risk of 30-day mortality following transcatheter aortic valve implantation (TAVI) using machine learning (ML) models. We commend the authors for developing two models based on TAVI pre-procedural (TRIMpre) and post-procedural (TRIMpost) variables; however, we would like to raise some concerns and discuss potential methodological challenges that might have influenced the results.

Model selection

In this study, the authors evaluated several ML models and utilized random forest models as the best-performing model, yet they did not report the performance metrics such as area under the curve (AUC) of alternative models.[1] The comparison between different ML models (such as logistic regression, support vector machines, decision trees, gradient boosting machines, and neural networks) can help identify potential biases and limitations in the developed predictive score.[2] Furthermore, it allows for the selection of a model that best suits the complexity of the data and would allow readers to better assess the robustness and generalizability of the selected model.

Class imbalance

The authors addressed the class imbalance within the study population by up-weighing the minority class (patients deceased within 30-day after TAVI) during the training phase. This strategy can enhance the model's sensitivity to detect rare events, though it may lead to overestimating event prevalence when predicting risks, particularly for patients outside of its training distribution. In addition, the authors tested an alternative approach without up-weighing of the minority class and enrolled the first 100 TAVI procedures per hospital that resulted in improved calibration but reduced performance metrics. Finally, they decided to use the up-weighing method and adhere to their proposed TRIM scores without reporting the sensitivity analysis results.

We believe that exploring alternative methods (such as oversampling) to address the class imbalance and discussing the trade-offs between calibration and classification performance would help readers understand the rationale behind them. Other approaches like the cost-sensitive learning methods can encourage the ML algorithm to focus on the minority class without requiring oversampling, which might be a better option in this setting by improving model performance while maintaining interpretability.[3]

Variable selection and feature importance

Many ML models summarize the impact of individual variables using feature importance metrics, which often present a ranking of the most influential variables for the fitted model. But, when deciding the number of predictors in the model, factors like data quality, overfitting prevention, and model interpretability must be considered alongside feature importance.[4] In this study, the authors developed and compared the performance of the aTRIMpost model using the top 5, 10, 20, and 25 important variables. However, it remains unclear how they ultimately decided on the number of features to include in the aTRIMpre model, and the performance test results are not reported.[1] The presented data show a significant decrease in feature importance between specific variable levels in the aTRIMpre model (−22 decrease in feature importance from variable level 15 to 16; supplementary material online, fig. S10).[1] This finding would suggest an optimal cut-off point for the number of predictors in the model, while maintaining similar performance using fewer predictors.

* Corresponding author. Tel: +98 21 8802 9254, Fax: +98 21 88029256, Email: kaveh_hosseini130@yahoo.com

Moreover, we noticed that some of the variables, like 'duration of fluoroscopy' and 'dose-area product', imply similar concepts (Fig. 3D).[1] The inclusion of such related variables raises questions about their individual added values, considering these features might be more reflective of the operator's skill and the complexity of the case rather than patient-specific risk factors. Also, several non-significant predictors such as 'height', 'peak to peak', 'left coronary artery', and 'month of admission' with a *P*-value of 1.0 were included in all models (supplementary material online, table S1).[1] We must highlight the importance of using the minimum number of variables in a model to reduce the risk of overfitting and enhance its applicability. For instance, the comparison between aTRIMpost and TRIMpost models exhibits similar positive likelihood ratios (PLRs) (TRIMpost PLR = 2.64, aTRIMpost PLR = 2.65) and only a marginal decrease in performance in the abridged version (delta AUC = 0.04), while requiring 130 (83.9%) fewer features. By selecting only the most relevant variables, the model can be substantially simplified and potentially become more robust when applied to unseen data.[5]

Lastly, it is striking that clinically significant predictors like baseline electrocardiogram and the incidence of pacemaker implementation after the procedure, which are known predictors of mortality and morbidity, were absent from the models.[6]

Importance of likelihood ratios in clinical practice

Likelihood ratios (LRs), both positive and negative, are extremely handy in clinical practice as they are not dependent on the underlying distribution of data. One of the many obstacles slowing down the adoption of ML applications in medicine is a poor performance on unseen data. Thus, when developing new algorithms and models, more emphasis should be put on minimizing input features and maximizing LRs rather than AUCs and other metrics. C-index or AUC primarily depicts the concordance of the predicted risks with the observed outcome but does not guarantee clinical usefulness.

Variable collinearity

Collinearity among input variables in a model can bias feature importance metrics and reduce model stability.[7,8] There are several approaches to assess collinearity in random forest models, including correlation analysis, variance inflation factor, permutation importance, etc.[7,8] In this study, the treatment of correlated variables or multicollinearity remains unclear. Moreover, the similar importance values in the feature importance plot (Fig. 3) and comparable patterns in partial dependence plots (Fig. 4) highly suggest collinearity between certain variables, such as 'duration of intervention', 'duration of fluoroscopy', and 'serum creatinine' or between 'peak aortic valve gradient (Pmax)' and 'weight'.[1] In light of these observations, the use of Explainable AI could elucidate the relationships between collinear variables in the model. However, this necessitates a thorough analysis to assess the model's variable collinearity.

In conclusion, while the study made a significant contribution to the development of ML models for risk assessment in patients undergoing TAVI, addressing the methodological concerns and potential biases mentioned above would further strengthen the study's findings and enhance their applicability to clinical practice.

Funding

No funding was used for the completion of this work.

Conflict of interest: None declared.

Data availability

The data underlying this article are available in the article.

References

1. Leha A, Huber C, Friede T, Bauer T, Beckmann A, Bekeredjian R, *et al.* Development and validation of explainable machine learning models for risk of mortality in transcatheter aortic valve implantation: TAVI risk machine scores. *Eur Heart J Digital Health* 2023;**4**:225–235.
2. Goldstein BA, Navar AM, Carter RE. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. *Eur Heart J* 2016;**38**:1805–1814.
3. Mienye ID, Sun Y. Performance analysis of cost-sensitive learning methods with application to imbalanced medical data. *Inform Med Unlocked* 2021;**25**:100690.
4. Speiser JL, Miller ME, Tooze J, Ip E. A comparison of random forest Variable selection methods for classification prediction modeling. *Expert Syst Appl* 2019;**134**:93–101.
5. Sanchez-Pinto LN, Venable LR, Fahrenbach J, Churpek MM. Comparison of variable selection methods for clinical predictive modeling. *Int J Med Inform* 2018;**116**:10–17.
6. Vicent L, Fernández-Cordón C, Nombela-Franco L, Escobar-Robledo LA, Ayesta A, Ariza Solé A, *et al.* Baseline ECG and prognosis after transcatheter aortic valve implantation: the role of interatrial block. *J Am Heart Assoc* 2020;**9**:e017624.
7. Strobl C, Boulesteix AL, Kneib T, Augustin T, Zeileis A. Conditional variable importance for random forests. *BMC Bioinformatics* 2008;**9**:307.
8. Cammarota C, Pinto A. Variable selection and importance in presence of high collinearity: an application to the prediction of lean body mass from multi-frequency bioelectrical impedance. *J Appl Stat* 2021;**48**:1644–1658.