









Challenges in developing and validating machine learning models for TAVI mortality risk prediction: reply

Andreas Leha^{1,2}, Cynthia Huber¹, Tim Friede ^{1,2}, Timm Bauer³,
Andreas Beckmann^{4,5}, Raffi Bekeredjian⁶, Sabine Bleiziffer ⁷, Eva Herrmann ^{8,9},
Helge Möllmann¹⁰, Thomas Walther ¹¹, Friedhelm Beyersdorf ^{12,13},
Christian Hamm^{14,15}, Arnaud Künzi¹⁶, Stephan Windecker ¹⁷, Stefan Stortecky¹⁷,
Ingo Kutschka ¹⁸, Gerd Hasenfuß^{2,19}, Stephan Ensminger ^{20,21},
Christian Frerker^{21,22}, and Tim Seidler ^{2,19,*}

¹Department of Medical Statistics, University Medical Center Göttingen, Humboldtallee 32, 37073 Göttingen, Germany; ²DZHK (German Center for Cardiovascular Research), Partner Site Göttingen, Robert-Koch str. 40, 37075 Göttingen, Germany; ³Department of Cardiology, Sana Klinikum Offenbach, Starkenburgring 66, 63069 Offenbach am Main, Germany; ⁴German Society for Thoracic and Cardiovascular Surgery, Langenbeck-Virchow-Haus, Luisenstraße 58/59, 10117 Berlin, Germany; ⁵Department for Cardiac and Pediatric Cardiac Surgery, Heart Center Duisburg, EVKLN, Gerrickstr. 21, 47137 Duisburg, Germany; ⁶Department of Cardiology, Robert-Bosch-Krankenhaus, Auerbachstraße 110, 70376 Stuttgart, Germany; ⁷Clinic for Thoracic and Cardiovascular Surgery, Heart and Diabetes Center Northrhine-Westphalia, Georgstr 11, 32545 Bad Oeynhausen, Germany; ⁸Goethe University Frankfurt, Department of Medicine, Institute of Biostatistics and Mathematical Modelling, Theodor-Stern-Kai 7, 60590 Frankfurt Main, Germany; ⁹DZHK (German Centre for Cardiovascular Research), Partner Site Rhine/Main, Theodor-Stern-Kai 7, 60590 Frankfurt Main, Germany; ¹⁰Department of Cardiology, St.-Johannes-Hospital Dortmund, Johannesstrasse 9-17, 44137 Dortmund, Germany; ¹¹Department of Cardiothoracic Surgery, University Hospital Frankfurt, Theodor-Stern-Kai 7, 60590 Frankfurt, Germany; ¹²Medical Faculty of the Albert-Ludwigs-University Freiburg, University Hospital Freiburg, Hugstetterstr. 55, 79106 Freiburg, Germany; ¹³Department of Cardiovascular Surgery, Heart Centre Freiburg University, Freiburg, Germany; ¹⁴Department of Cardiology and Angiology, University Hospital Gießen, Klinikstr. 33, 35392 Gießen, Germany; ¹⁵Department of Cardiology, Kerckhoff Heart and Thorax Center, Benckestraße 2-8, D-61231 Bad Nauheim, Germany; ¹⁶CTU Bern, University of Bern, Mittelstrasse 43, 3012 Bern, Switzerland; ¹⁷Department of Cardiology, Inselspital, Bern University Hospital, University of Bern, 3010 Bern, Switzerland; ¹⁸Clinic for Cardiothoracic and Vascular Surgery/Heart Center, University Medical Center Göttingen, Robert-Koch Str. 40, 37075 Göttingen, Germany; ¹⁹Clinic for Cardiology and Pulmonology, Heart Center, University Medical Center Göttingen, Robert-Koch Str. 40, 37075 Göttingen, Germany; ²⁰Department of Cardiac and Thoracic Vascular Surgery, University Heart Center Lübeck, Ratzeburger Allee 160, 23538 Lübeck, Germany; ²¹DZHK (German Centre for Cardiovascular Research), partner site Hamburg/Kiel/Lübeck, Lübeck, Germany; and ²²Department of Cardiology, University Heart Center Lübeck, Ratzeburger Allee 160, 23538 Lübeck, Germany

Received 2 October 2023; accepted 4 October 2023; online publish-ahead-of-print 8 November 2023

Commentary article to: ‘Challenges in developing and validating machine learning models for transcatheter aortic valve implantation mortality risk prediction’, by S. Kazemian *et al.*, <https://doi.org/10.1093/ehjdh/ztd059>.

We welcome a discussion of our article¹ and thank Kazemian *et al.*² for their constructive comments on how to expand the analyses and in particular the reporting. At the same time, we are relieved that no points were brought forward that would potentially draw our investigations into question. Therefore, we stand by our conclusions and recommend the transcatheter aortic valve implantation (TAVI) risk machine (TRIM) scores for application. In the following, we address all comments by Kazemian *et al.* in order.

Model selection

We agree that giving details also on the process on how researchers arrive at their final results is good practise and enhances transparency. Giving more details on the alternative machine learning (ML) models might be of interest in particular for the ML expert reader.

Listing the performance measures of the alternative ML models would, however, undoubtedly lead to the next questions on how exactly these other ML models were trained. The description of the exact training of all models would again take considerable space—especially for deep learning type models—and is out of scope for this article which has its focus on the application and is already quite demanding on the reader regarding methodology. We did not mean to compare various methods but rather to provide appropriate risk models.

In our view, more methodology focused types of publications, like systematic reviews³ or neutral comparison studies,⁴ are the place for more in-depth discussions and comparisons of ML models, and we urge the community for more research in these directions.

It is not surprising that different ML models show different performance, as there is no ML model that performs best on all applications. This is well-known in the field and has been acknowledged in the review paper⁵ cited by Kazemian *et al.*

As others have noted that there is no best approach for all data problems. The various techniques differ in their approaches as they aim to solve different data complexities. Therefore, the ‘best’ algorithm will depend on the specific data problem at hand.

* Corresponding author. Tel: +49 (0) 551/39 63907, Fax: +49 (0) 551/39 63906, Email: tim.seidler@med.uni-goettingen.de

© The Author(s) 2023. Published by Oxford University Press on behalf of the European Society of Cardiology.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Importantly, the inclusion of more performance measures for alternative ML models would not enhance the reader's ability to 'assess the robustness and generalizability of the selected model' much. Instead, robustness and generalizability have to be assessed separately for each model via careful validation. This is exactly what we did for the selected random forest model.

Class imbalance

Kazemian *et al.* address the point in our article that the proposed ML score overestimates the probability of events and is, thus, not interpretable as patient risk, i.e. not well calibrated. In the article, we describe the up-weighting of the minority class as one contributing factor and present a model trained without such up-weighting, which is better calibrated but shows numerically lower performance. Kazemian *et al.* now suggest to use alternative methods such as oversampling or cost-sensitive ML models to deal with class imbalance.

While cost-sensitive modelling might improve calibration without negative impact on performance, Kazemian *et al.* fail to consider that the up-weighting in random forest models is basically a form of random oversampling, as the weight simply corresponds to the probability of patients to be sampled within the bagging of the random forest. While there are more sophisticated oversampling techniques such as SMOTE and we do not know the impact on the classification of these data (no free lunch), there are papers that show that such more sophisticated methods often lead to inferior performance compared with simpler methods:

One of the most important conclusions that can be drawn from these experiments is the inferior performance of the 'intelligent' sampling techniques, SM [SMOTE], BSM, WE, OSS, and CBOS.⁶

Kazemian *et al.* further state 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'.

As Kazemian *et al.* are aware, our paper already includes a discussion on the trade-off, see 'The selection of patients with events with increased probability during training, which we implemented to address the class imbalance, leads to an overestimation of the prevalence of events. When we omitted the up-weighting of the minority and counteracted the expected time-effect by disregarding the first 100 TAVI interventions per hospital, a more favourable calibration was achieved'.

Any further or more in-depth discussion on dependencies between calibration and classification performance would be a technical discussion involving how and why random forests might be affected by class imbalance and to which extent the area under the curve (AUC) as a measure of performance might be affected by class imbalance. Such discussion as well as the exploration of other methods to handle class imbalance is again not in the focus of the paper but would be much better suited for a more methodology-focused article.

Variable selection and feature importance

Kazemian *et al.* claim that the performance test results are not reported. Here, they fail to see that Supplementary material online, Figure S6 shows the performance results. Kazemian *et al.* also suggest that cutting at the top 15 variables would lead to a simpler model 'maintaining similar performance using fewer predictors'. We see, however, that between using 20 and 25 variables, the performance (measured as AUC) increases from 0.716 to 0.74. As we write in our paper, 'RF performance continued to improve with more features and reached its maximum with the entire 155 feature-set'. We can conclude that single feature importance might seem small, but they can still show cumulative

effects on performance. Importantly, we present feature rich models (in addition to our presentation of models with only few variables) to address the possibility of automated data transfer from hospital information systems, not manual entry.

Furthermore, Kazemian *et al.* advocate for simpler models using fewer variables for enhanced applicability and to reduce the risk of overfitting, so that they 'potentially become more robust when applied to unseen data'. This is true in principle, and we fully agree that simpler models are preferable for these reasons as well as for other reasons like easier interpretability. However, Kazemian *et al.* fail to see that if the larger models were less robust in this application, they would show lower performance on the external test data from SwissTAVI, but the opposite is the case. So, the non-abridged versions of the scores show higher robustness. We also dispute the claim that the larger amount of variables reduces applicability. While again this is true in principle, we show that the scores remain applicable even with large amounts of data missing. Again, the fraction of missing variables was largest for the full score, and yet its performance stayed superior.

Kazemian *et al.* also state that variables with *P*-values of 1.0 are included in all models. We are not surprised that some variables do not show any association when tested univariably between the two groups of patients yet get high importance values from the random forest models (and thus are selected also for the abridged versions of the scores). The ability of the models to unravel heterogeneous effects and non-linearities is one of the reasons why ML models are applied as nicely summarized in⁵:

This algorithm [classification and regression tree] is typically able to handle all three challenges of non-linearities, heterogeneous effects, and many predictors.

An excellent example for the observation that *P* might not be relevant to assess feature importance is 'peak to peak' pressure, which shows a U-shaped curve describing its influence on the prediction in our model, which is something that is not easily seen in a univariable test.

Lastly, Kazemian *et al.* note a 'striking' lack of known predictors, such as baseline electrocardiogram and incidence of pacemaker implementation. Notably, the literature is not in uniform support of pacemaker dependency or conduction abnormalities as independent predictors of outcomes (reviewed in Sammour *et al.*⁷). Moreover, conduction abnormalities like left bundle branch block frequently resolve within days after TAVI and are certainly not ideal for a decision support model before and directly after TAVI. The models were trained on registry data, and all data available at the decision time points were used to train the models.

Our models predict outcomes either before (TRIMpre) or directly after (TRIMpost) the procedure to reflect time of decision-making in clinical practice. (It is useless to provide decision support regarding early discharge based on data derived at late discharge.) New pacemaker dependency after TAVI is certainly not known pre-operatively and in many cases also not immediately after the intervention. (Note that if patients had a pacemaker before TAVI, this was part of the training.) Thus, we deliberately did not include this variable in the set of potential predictors, just as the post-interventional complications.

Importance of likelihood ratios in clinical practice

Kazemian *et al.* argue that the AUC primarily represents the concordance of the predicted risks with the observed outcome and is, thus, less useful than likelihood ratios.

We agree that likelihood ratios are an excellent addition to reporting sensitivities and specificities, as they are easy to interpret and largely independent of the prevalence, so that they generalize more easily to new

data. In addition, likelihood ratios can also be used to compare the performance of different diagnostic tests (or different ML models).^{8,9}

However, likelihood ratios in their basic form are based on binary classifiers. If an ML model predicts risk scores, one has to choose a cut-off for, say, high-risk vs. low-risk in order to calculate sensitivity, specificity, and the likelihood ratios.

While in the end, a decision must be taken (e.g. discharge early or not) and this decision should be taken by the physician and not by the model and the model should merely serve as decision support.¹⁰ For such decision support, the confidence of the model is important, as a predicted risk merely above a threshold is a vastly different result than a predicted risk of 1. The model should, therefore, return the risk or risk score to be interpreted by the physician. Thus, we argue that performance measures that necessitate a binary decision of the model are not measuring the full picture.

The AUC as a measure of discrimination performance does not require the specification of cut-offs as it summarizes over all possible cut-offs. For likelihood ratios, there exists the concept of multi-level likelihood ratios that allow for more than a single cut-point. But then, multi-level likelihood ratios become unwieldy as they do not provide a single measure anymore, and comparisons between several classifiers based on multi-level likelihood ratios are not straight forward and typically assume matching levels between the compared models.^{8,9}

Furthermore, Kazemian *et al.* state: 'One of the many obstacles slowing down the adoption of ML applications in medicine is poor performance on unseen data.' We fully agree with this statement. In our view, the lack of validation on unseen and at best external data is the main contributing factor here. In our article, we, thus, reported the performance on external test data providing an unbiased assessment of the proposed risk scores—despite many missing values.

Variable collinearity

Kazemian *et al.* express concerns about including correlated variables as that can bias feature importance metrics and reduce model stability.

It is true that the variable importance measures can be biased by correlated variables depending on the degree of correlation, the size of the groups of correlated variables, and then the number of previously selected splitting variables (the *mtry* parameter) used in

the training (see Gregorutti *et al.*¹¹ and Strobl *et al.*¹² for discussions on permutation importance). As we have not specifically dealt with correlated variables (removing any correlated variables from the models, using principle component analysis on clusters, or similar) and are also not reporting conditional importances suggested in,¹² the variable importance estimates for correlated variables might, indeed, be inflated.

The predictive performance of the random forest, however, should not be affected by the inclusion of correlated variables.

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 Digit Health* 2023; **4**:225–235.
2. Kazemian S, Issayi M, Hosseini K. Challenges in developing and validating machine learning models for transcatheter aortic valve implantation mortality risk prediction. *Eur Heart J Digit Health* 2024; **5**:1–2.
3. Friedrich S, Groß S, König IR, Engelhardt S, Bahls M, Heinz J, *et al.* Applications of artificial intelligence/machine learning approaches in cardiovascular medicine: a systematic review with recommendations. *Eur Heart J Digit Health* 2021; **2**:424–436.
4. Boulesteix AL, Lauer S, Eugster MJA. A plea for neutral comparison studies in computational sciences. *PLoS One* 2013; **8**:e61562.
5. Goldstein BA, Navar AM, Carter RE. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. *Eur Heart J* 2017; **38**:1805–1814.
6. Hulse JV, Khoshgoftaar TM, Napolitano A. Proceedings of the 24th International Conference on Machine Learning, Corvallis, Oregon, USA. 2007.
7. Sammour Y, Krishnaswamy A, Kumar A, Puri R, Tarakji KG, Bazarbashi N, *et al.* Incidence, predictors, and implications of permanent pacemaker requirement after transcatheter aortic valve replacement. *JACC Cardiovasc Interv* 2021; **14**:115–134.
8. Luts J, Nofuentes JAR, de Dios Luna del Castillo J, Huffel SV. Asymptotic hypothesis test to compare likelihood ratios of multiple diagnostic tests in unpaired designs. *J Stat Plan Inference* 2011; **141**:3578–3594.
9. Leisenring W, Pepe MS. Regression modelling of diagnostic likelihood ratios for the evaluation of medical diagnostic tests. *Biometrics* 1998; **54**:444–452.
10. Pate A, Emsley R, Ashcroft DM, Brown B, van Staa T. The uncertainty with using risk prediction models for individual decision making: an exemplar cohort study examining the prediction of cardiovascular disease in English primary care. *BMC Med* 2019; **17**:134.
11. Gregorutti B, Michel B, Saint-Pierre P. Correlation and variable importance in random forests. *Stat Comput* 2016; **27**:659–678.
12. Strobl C, Boulesteix AL, Kneib T, Augustin T, Zeileis A. Conditional variable importance for random forests. *BMC Bioinformatics* 2008; **9**:307.