The Combination of Supervised and Unsupervised Learning based Risk Stratification and Phenotyping in Pulmonary Arterial Hypertension - a Long-term Retrospective Multicenter Trial

PULM-D-22-00488, point-to-point reply

Innsbruck PAH registry

2023-03-27

# Editor

We would like to thank the Editor and Team of BMC Pulmonary Medicine for excellent editorial work. Independently of the Relievers’ Issues, we have introduced few minor changes to the R pipeline and the manuscript to accommodate the latest versions of analysis software and improve the readability.

* The R code analysis pipeline was available at <https://github.com/PiotrTymoszuk/PAH-biomarker> was adopted for the R version 4.2.3 and the latest versions of analysis packages. This had not effect on analysis results.
* The manuscript text was shortened, e.g. by removal of phrases wording and suppression of suggestive wording (‘notably’, ‘importantly’) to accommodate changes requested be the Reviewers.
* **Figure 2A**: variables of the Elastic Net signature were ordered by the HR values.
* **Figure 2B**: the Kaplan Meier plots depicting the actual and fitted survival curves were removed since we considered them not particularly informative. In the revised manuscript, we present Kaplan Meier survival curves for tertiles of the Elastic Net signature at this place, which, we feel are more important for visual interpretation of the modeling results. Extensive numeric characteristic of goodness-of-fit of the Elastic Net model and the established risk assessment tools is presented now in **Supplementary Table S4**.
* **Figure 4B**: in the revised manuscript, violin plots of the clustering factors are appended with p values for differences between the clusters. We consider this solution helpful for instant interpretation of differences between the clusters.
* **Supplementary Figure S5**: newer versions of our standard clustering tool, the development package [*clustTools*](https://github.com/PiotrTymoszuk/clustTools), allow now for robust testing of permutation importance of clustering factors for multiple random re-shuffles. For this reason, we analyzed clustering factor importance for n = 100 random runs. This approach is considered to improve reproducibility.
* **Supplementary Tables**: we have reduced the number of footnotes and repeated wording ‘median’ and ‘range’. We describe instead the format of variables in the table captions.

\***Supplementary Methods**: were adapted to the changes in R markup. We also include few more details on R functions used in data analysis, in line with our recent publications.

# Reviewer 1

We would like to thank the Reviewer for careful lecture and valuable feedback.

## Issue 1

Considering that cluster analysis is a kind of unsupervised learning in machine learning, the title of the paper is suggested to be “The Combination of Supervised and Unsupervised Learning based Risk Stratification and Phenotyping in Pulmonary Arterial Hypertension - a Long-term Retrospective Multicenter Trial”. Additionally, in this paper, “PAM algorithm and cluster analysis” is used instead of “machine learning and cluster analysis”.

## Response 1

We thank the Reviewer for the notion and adapted the title as suggested. Given the plethora of machine learning algorithms we now describe more precisely the key methods as ‘Elastic Net modeling/regression’ or ‘PAM clustering’ throughout the text and suppressed the use of the general term ‘machine learning’.

## Issue 2

When increasing data and index, or applying another cluster method, are these two clusters still suitable? Maybe there are three clusters consistent with the three risk levels of PAH (low, medium and high)?

## Response 2

We are grateful for this interesting point. Our intention behind the clustering analysis was identification of clinically interesting risk patterns based on factors found to be associated with overall survival in Elastic Net regression rather than recapitulation of the ‘lo/intermediate/high’ risk classification scheme followed by most PAH risk scales.

To clarify the cluster number issue, we present now the optimal cluster number based on the bend of WSS curve (within-cluster sum of squares) and maximum of mean silhouette statistic for each of the algorithms compared with the PAM/cosine procedure in **Supplementary Figure S4A**. Interestingly, the PAM/cosine algorithm was found more stable (10-fold cross-validation) [1] and more informative (explained clustering variance) than its hierarchical clustering or k-means competitors with k = 3 clusters. Concerning the algorithm of choice, the solution with k = 2 clusters was found optimal as investigated with three independent methods: the WSS curve, silhouette statistic [2] and cross-validation [1] (**Supplementary Figure S4B**). In addition, a visual inspection of UMAP layouts [3, 4] which we present in the revised manuscript instead of PCA scores speaks for two clearly separated clusters in each of the cohorts (**Figure 4A**). Following the Reviewer’s suggestion, we have investigated a larger pooled data set with the pooled Innsbruck and Linz/Vienna cohorts. As presented in **Reviewer Figure S1**, the two cluster PAM/cosine distance algorithm was found again the most stable (accuracy = 0.95) and acceptably informative (explained variance = 0.45) clustering solution. Collectively, we are pretty confident that the study participants can be reproducibly assigned to two subsets based on age, SMWD, RDW, cardiac index, PVR, NT-pro-BNP and RAA. Whether this classification applies to other larger cohorts, remains to be investigated. We discuss the results of comparison of clustering algorithms and the extended cluster number finding in **Results** of the revised manuscript.

# Reviewer 2

We thank the Reviewer for the thorough inspection of the manuscript and the feedback.

## Issue 3

In this manuscript, Sonnweber and colleagues present an analysis of prognostic indicators in pulmonary arterial hypertension (PAH) patients from 3 cohorts, grouped into two analyses. They use machine learning and clustering of 19 established clinical/hospital laboratory values to build a model containing age, six-minute walking distance, red blood cell distribution width, cardiac index, pulmonary vascular resistance, N-terminal pro-brain natriuretic peptide and right atrial area. None of these are surprising variables although RDW is often overlooked. This model is validated in the second analysis of Linz/Vienna patients. The clustering defines two clusters, one higher risk and older and the other with typical lower risk profiles. The clusters are also associated with established clinical risk scores, which include some of the same variables. The study is a proof of concept rather than definitive in of itself and provides an example of how these methods could be used with richer datasets, for example proteomics, as suggested by the authors.

## Response 3

As outlined by the reviewer, the presented analysis, which includes 183 PAH patients, presents a proof of concept analysis and further evaluation in larger cohorts is warranted. One aim of the study was to test applicability of established machine learning algorithms like Elastic Net regression [5] or unsupervised clustering [6] in development of efficient and reproducible risk models. We herein focused on currently broadly available parameters for PAH risk assessment and their highly predictive combinations, rather than on identification of entirely novel risk factors. We totally agree with the reviewer that several other parameters (including serum markers of inflammation or vascular damage) [7] may improve the accuracy of survival prediction in PAH and may be an included in our risk modeling approach in the future. Intriguingly, to date only a few of the known risk factors for PAH mortality are available in clinical PAH registers and even fewer are used in clinical practice. Accordingly, the use of complex PAH risk scores, such as the 14-parameter REVEAL 2.0 [8], demonstrate a good predictive performance but are challenged by a resource-consuming process with limits their applicability in clinical practice. Thus, the herein presented risk assessment approach aims to identify the most robust, reproducible and easily applicable risk parameter combination, the Elastic Net signature, including currently available risk markers. Whereas this approach deliberately lacks the novelty of new risk parameters, it may provide an easy path to improve PAH risk assessment in clinical practice without the need to spend significant additional resources. In line with the comments of Reviewers 2 and 3, we extended Discussion of the revised manuscript, which now stresses the aim of the study more clearly and highlights the usefulness of new additional parameters for PAH risk assessment.

## Issue 4

There is no attempt to demonstrate that the model/clusters derived add anything in terms of performance to established measures/models for example the risk equations/models mentioned, in a head-to-head or combinatorial model. Again, as the authors state, this is a proof-of-concept rather than a novel model with suggested utility.

## Response 4

This is an interesting point. Intriguingly, in the newly developed Elastic Net signature correlated with risk classification by established tools (mRASP, COMPERA, SPAHR, FPHR 3p and FPHR 4p) to a lesser extent (Spearman’s rho: 0.57 – 0.79) than the established risk scales with each other (rho: 0.61 – 0.9) (**Supplementary Figure S2**). This may indicate that the Elastic Net model may indeed ‘cover’ patient subsets misclassified by the existing risk assessment tools.

In the revised manuscript, we have compared performance of the newly developed Elastic Net signature with the established risk assessment tools: mRASP, COMPERA, SPAHR, FPHR 3p and FPHR 4p. Additionally, we have developed an ensemble survival model with the Ridge Cox regression procedure [5] employing the established risk assessment tools as explanatory variables. Of note, high risk strata classification with SPAHR and mRASP were the strongest unfavorable covariates of the ensemble model (**Supplementary Figure S3A**).

Performance of those models was compared with three metrics: Harrell’s concordance index (C) [9], R2 [10] and integrated Brier score [[11]. As shown in **Figure 3** and **Supplementary Table S4**, the Elastic Net signature outcompeted all established comparator risk assessment tools in terms of goodness-of-fit (C, IBS) and explained variance (). The ensemble of those popular risk assessment scales and the Elastic Net signature had comparable survival prediction performance in the IBK cohort, but the ensemble model performed poorly in the validation LZ/W collective. This indicates that an optimal combination of easily available prognostic factors may measurably improve the death risk estimation. Following the Reviewer’s suggestion, we present and discuss the comparison results in the revised manuscript.

Finally, we constructed another ensemble of the mRASP, COMPERA, SPAHR, FPHR 3p and FPHR 4p in the Innsbruck cohort with the Survival Random Forest algorithm which is assumed to handle highly inter-correlated explanatory factors and interactions between explanatory factors more robustly than regression methods [12, 13]. Interestingly, such procedure generated a model which relied on three-parameter FPHR, mRASP and COMPERA to the highest extent (**Figure 2**). Interestingly, prediction accuracy of this more sophisticated tool was still worse than of the Elastic Net model. Collectively, the data for established PAH risk assessment tools and their ensembles show their performance limits and underline the need for novel statistical approaches to modeling survival in PAH.

## Issue 5

In methods, median centering and “To account for non-linear associations of numeric independent variables, both first and second-order terms were included in the model” is repeated, and can be simplified.

## Response 5

We thank the Reviewer for the suggestion and adapted **Methods** of the revised manuscript. In addition, we redacted the revised text and removed repeated phrases for the sake of readability.

## Issue 6

To include n = 8 CTD-PAH patients in a paper of n~200 patients seems unnecessary/underpowered to assess potential differential effects/utility.

## Response 6

We agree with the reviewer that the CTD-PAH group is not large enough for subgroup analysis. Still, risk assessment is well established for CTD-PAH, thus we think it is appropriate to include CTD-PAH in a PAH risk assessment evaluation.

## Issue 7

RDW has been previously identified as a good prognostic marker in PAH but this is not clear in the discussion e.g. Heart. 2011 Jul;97(13):1054-60. doi: 10.1136/hrt.2011.224857.

## Response 7

We thank the reviewer for this notion and included a statement and the reference [14] in **Discussion** of the manuscript.

## Issue 8

Fig 1 - LZ/W says parameter missing n = 30 but only 12 lost?

## Response 8

We apologize for the typo in **Figure 1** and thank the Reviewer for careful lecture. The number of participants excluded due to data missingness (n = 30) is correct, but the final number was wrong. The final size of Linz/Vienna cohort is n = 83 as reported in the text. **Figure 1** was corrected accordingly.

## Issue 9

Some parts of the supplemental tables are hard to read because the words median/IQR/range are repeated so much - it would be easier if the data were shown on their own, structured in a clear way such that they didn’t need to be re-labeled on each line.

## Response 10

We thank the Reviewer for the feedback. In the revised **Supplementary Material**, we have paid particular attention to readability of Supplementary Tables. As suggested, we removed the repeated wording ‘median’ and ‘range’ and describe instead the format of numeric and categorical variables in table captions. Additionally, we have compacted table footnotes and included in **Supplementary Table S1** only variables used in the data analysis. Independently of Reviewer’s suggestions, we adapted the analysis pipeline to the latest version of R (4.2.3) and R packages used for data transformation and analysis. This had no effect on the analysis results, but shortened and improved readability of the R markup. We introduced also minor changes to manuscript Figures such as arranging of HR values in **Figure 2A** or displaying of p values in the plot axis in **Figure 4B**, which, in our opinion, will aid result interpretation.

# Reviewer 3

We thank the Reviewer for the careful lecture and feedback.

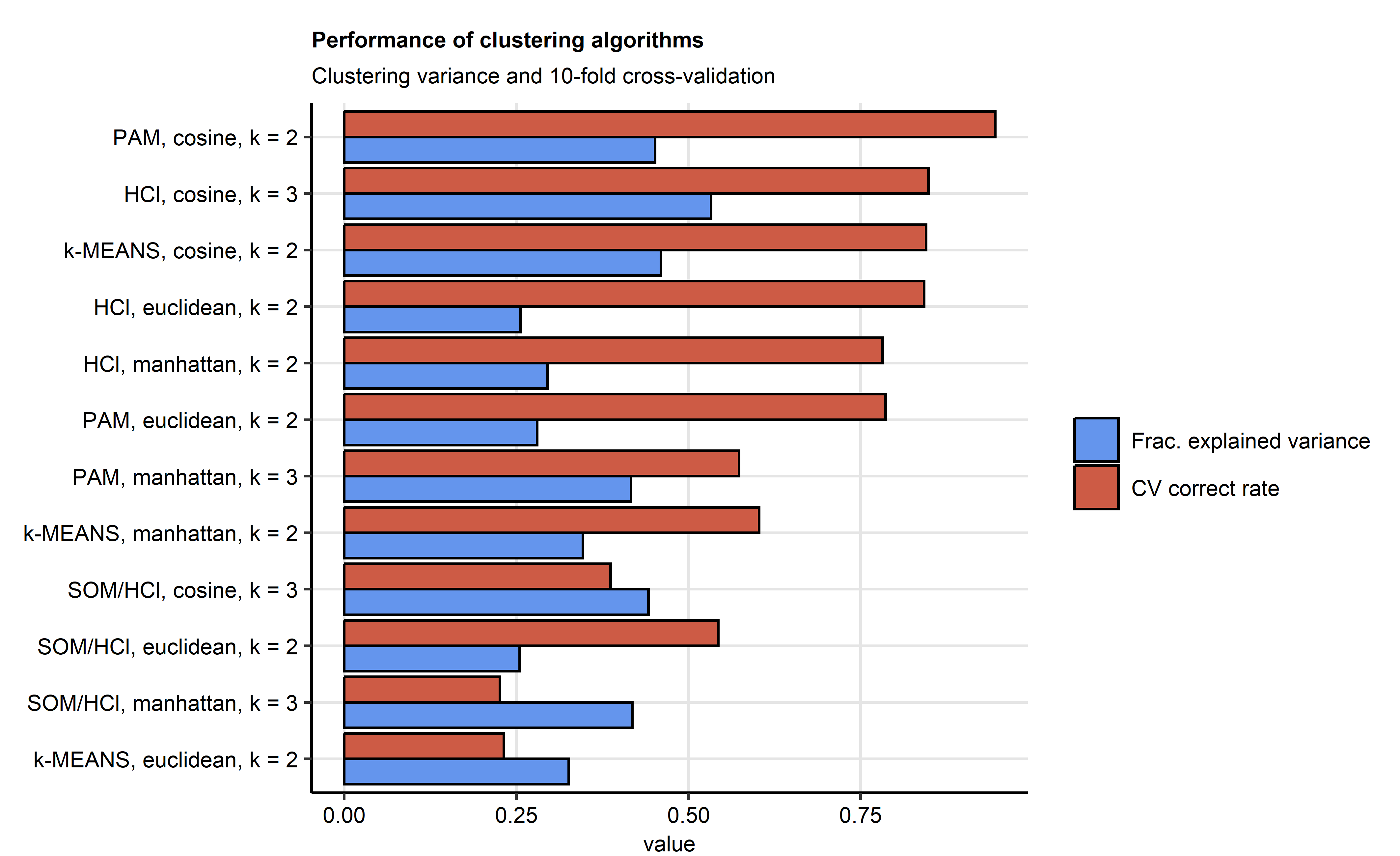
## Issue 11

In this retrospective, multicenter, observational two-cohort study, 183 PAH patients were analyzed. It has been shown that seven parameters, including age, six-minute walking distance, red blood cell distribution width, cardiac index, pulmonary vascular resistance, N-terminal pro-brain natriuretic peptide, and right atrial area, determine mortality. It was concluded that machine learning and clustering algorithms are important tools for automatic mortality risk estimation. With this machine learning strategy, only the potential benefit of red blood cell distribution width has been mentioned. However, factors contributing to unfavorable prognosis such as advanced age at the time of diagnosis, poor cardiac output, high pulmonary vascular resistance, poor six-minute walk test performance have not been adequately discussed. The results of the study will contribute to the literature. If the discussion is expanded in an appropriate manner, it may be published.

## Response 11

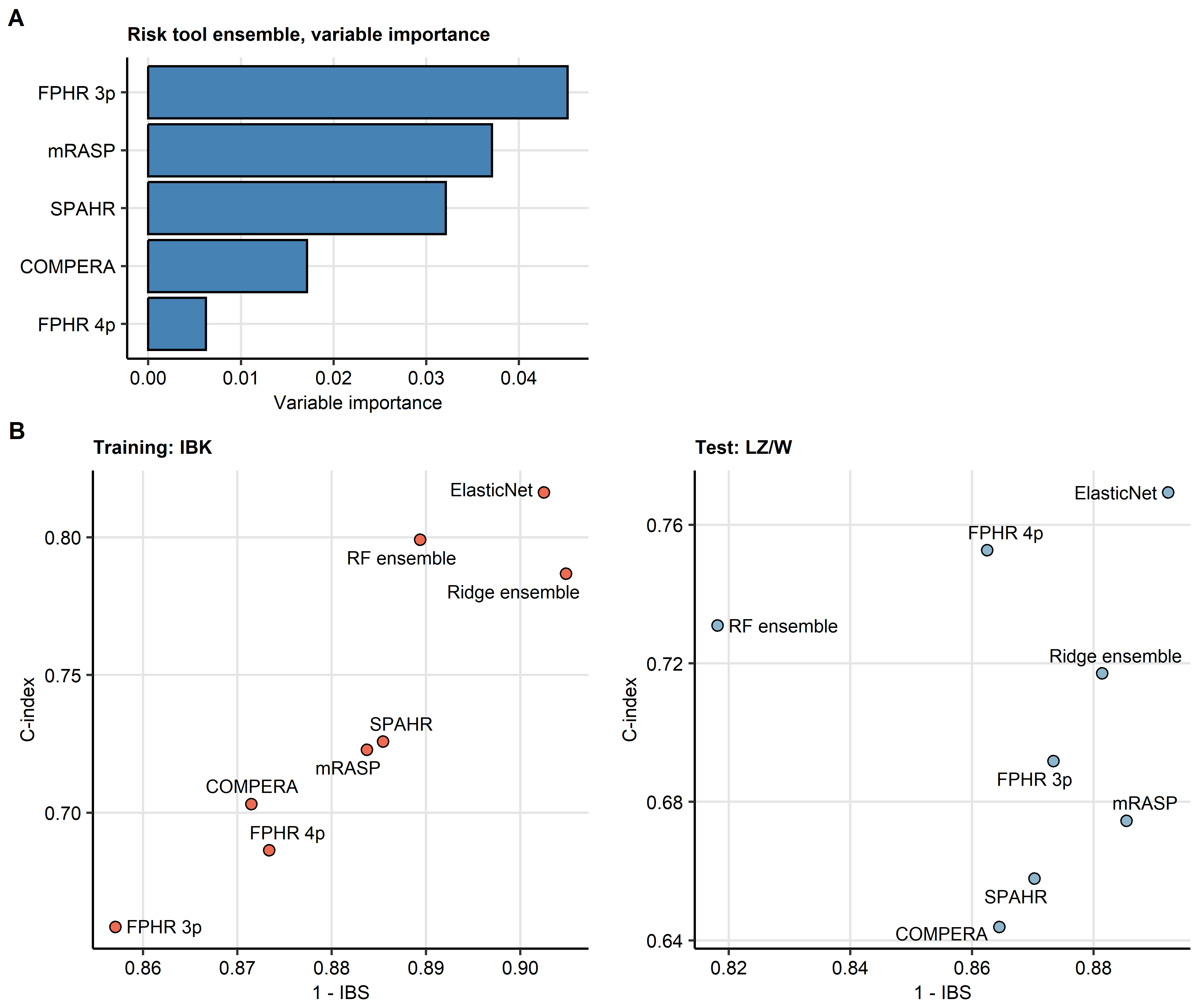
We agree with the Reviewer that the manuscript can be improved by expanding **Discussion** as suggested. Thus, we included a new paragraph, which addresses the role of unfavorable parameters for PAH mortality risk.

# Figures



**Reviewer Figure R1. Clustering analysis with the pooled IBK and LZ/W data set.**

*Clustering of the pooled data set with participants of the Innsbruck (IBK) and Linz/Vienna cohorts in respect to the survival-associated factors identified by elastic-net modeling (Figure 2A) was investigated by several algorithms (PAM: partitioning around medoids, HCl: hierarchical clustering, k-MEANS and SOM/HCl: combined self-organizing map/hierarchical clustering). The optimal cluster number k was determined by analysis of the curve of within-cluster sum of squares and maximal silhouette statistic. Explanatory performance of the algorithms was assessed by the fraction of ‘explained’ clustering variance (ratio of between-cluster sum of squares to total sum of squares) and correct cluster assignment rate in 10-fold cross-validation (CV). Statistic values are presented in bar plots. Algorithm, distance metric and k number of clusters are indicated in the Y axis.*



**Reviewer Figure R2. Development of a Random Forest ensemble of PAH risk assessment tools and its performance at predicting overall survival.**

*Elastic Net signature was developed as presented in Figure 2. The ensemble models of the Elastic Net score and established PAH risk assessment tools (FPHR 3p: French Pulmonary Hypertension Registry 3 parameter score, FPHR 4p: French Pulmonary Hypertension Registry 4 parameter score, COMPERA: Comparative, Prospective Registry of Newly Initiated Therapies for Pulmonary Hypertension score, mRASP: modified Risk Assessment Score of PAH.) was established with the LASSO Cox regression (Supplementary Figure S3) and survival Random Forest algorithms (number of trees: 1000, mtry = 3, splitting rule: logrankscore, minimal node size: 6).*

*(A) Permutation importance of explanatory variables of the RF ensemble model.*

*(B) Predictive performance of the Elastic Net signature, LASSO ensemble, Random Forest (RF) ensemble and single PAH risk scores at predicting overall survival was assessed by concordance index (C-index) and integrated Brier score (IBS). C-indexes and IBS for the risk assessment tools in the Innsbruck (IBK) and Linz/Vienna (LZ/W) cohorts are displayed in scatter plots.*

# References

1. Lange T, Roth V, Braun ML, Buhmann JM. [Stability-based validation of clustering solutions](https://doi.org/10.1162/089976604773717621). Neural Computation. 2004;16:1299–323.

2. Rousseeuw PJ. [Silhouettes: A graphical aid to the interpretation and validation of cluster analysis](https://doi.org/10.1016/0377-0427(87)90125-7). Journal of Computational and Applied Mathematics. 1987;20 C:53–65.

3. Konopka T. [umap: Uniform Manifold Approximation and Projection](https://cran.r-project.org/web/packages/umap/index.html). 2022.

4. McInnes L, Healy J, Melville J. [UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction](https://arxiv.org/abs/1802.03426v3 http://arxiv.org/abs/1802.03426). 2018.

5. Simon N, Friedman J, Hastie T, Tibshirani R. [Regularization Paths for Cox’s Proportional Hazards Model via Coordinate Descent](https://doi.org/10.18637/JSS.V039.I05). Journal of Statistical Software. 2011;39:1–3.

6. Schubert E, Rousseeuw PJ. [Faster k-Medoids Clustering: Improving the PAM, CLARA, and CLARANS Algorithms](https://doi.org/10.1007/978-3-030-32047-8_16). In: Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics). Springer; 2019. p. 171–87.

7. Kanwar MK, Gomberg-Maitland M, Hoeper M, Pausch C, Pittow D, Strange G, et al. [Risk stratification in pulmonary arterial hypertension using Bayesian analysis](https://doi.org/10.1183/13993003.00008-2020). The European respiratory journal. 2020;56.

8. Benza RL, Gomberg-Maitland M, Elliott CG, Farber HW, Foreman AJ, Frost AE, et al. [Predicting Survival in Patients With Pulmonary Arterial Hypertension: The REVEAL Risk Score Calculator 2.0 and Comparison With ESC/ERS-Based Risk Assessment Strategies](https://doi.org/10.1016/J.CHEST.2019.02.004). Chest. 2019;156:323–37.

9. Harrell FE, Lee KL, Mark DB. [Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors](https://doi.org/10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4). Statistics in Medicine. 1996;15:361–87.

10. Dardis C. [survMisc: Miscellaneous Functions for Survival Data](https://cran.r-project.org/web/packages/survMisc/index.html). 2022.

11. Graf E, Schmoor C, Sauerbrei W, Schumacher M. [Assessment and comparison of prognostic classification schemes for survival data.](https://doi.org/10.1002/(sici)1097-0258(19990915/30)18:17/18<2529::aid-sim274>3.0.co;2-5) Statistics in Medicine. 1999;18:2529–45.

12. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. [Random survival forests](https://doi.org/10.1214/08-AOAS169). https://doiorg/101214/08-AOAS169. 2008;2:841–60.

13. Breiman L. [Random forests](https://doi.org/10.1023/A:1010933404324). Machine Learning. 2001;45:5–32.

14. Rhodes CJ, Wharton J, Howard LS, Gibbs JSR, Wilkins MR. [Red cell distribution width outperforms other potential circulating biomarkers in predicting survival in idiopathic pulmonary arterial hypertension](https://doi.org/10.1136/HRT.2011.224857). Heart (British Cardiac Society). 2011;97:1054–60.