


RESEARCH ARTICLE

Open Access



The combination of supervised and unsupervised learning based risk stratification and phenotyping in pulmonary arterial hypertension—a long-term retrospective multicenter trial

Thomas Sonnweber^{1*} , Piotr Tymoszek², Regina Steringer-Mascherbauer³, Elisabeth Sigmund³, Stephanie Porod-Schneiderbauer³, Lisa Kohlbacher⁴, Igor Theurl¹, Irene Lang⁴, Günter Weiss¹ and Judith Löffler-Ragg¹

Abstract

Background Accurate risk stratification in pulmonary arterial hypertension (PAH), a devastating cardiopulmonary disease, is essential to guide successful therapy. Machine learning may improve risk management and harness clinical variability in PAH.

Methods We conducted a long-term retrospective observational study (median follow-up: 67 months) including 183 PAH patients from three Austrian PAH expert centers. Clinical, cardiopulmonary function, laboratory, imaging, and hemodynamic parameters were assessed. Cox proportional hazard Elastic Net and partitioning around medoid clustering were applied to establish a multi-parameter PAH mortality risk signature and investigate PAH phenotypes.

Results Seven parameters identified by Elastic Net modeling, namely 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, constituted a highly predictive mortality risk signature (training cohort: concordance index = 0.82 [95%CI: 0.75 – 0.89], test cohort: 0.77 [0.66 – 0.88]). The Elastic Net signature demonstrated superior prognostic accuracy as compared with five established risk scores. The signature factors defined two clusters of PAH patients with distinct risk profiles. The high-risk/poor prognosis cluster was characterized by advanced age at diagnosis, poor cardiac output, increased red cell distribution width, higher pulmonary vascular resistance, and a poor six-minute walking test performance.

Conclusion Supervised and unsupervised learning algorithms such as Elastic Net regression and medoid clustering are powerful tools for automated mortality risk prediction and clinical phenotyping in PAH.

Keywords Pulmonary arterial hypertension, Risk assessment, Biomarkers, Mortality, Right-heart failure, Atypical pulmonary arterial hypertension

*Correspondence:

Thomas Sonnweber

Thomas.Sonnweber@i-med.ac.at; judith.loeffler@i-med.ac.at

Full list of author information is available at the end of the article



© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Background

Pulmonary arterial hypertension (PAH) is a rare disease with a detrimental long-term outcome [1–4]. Hemodynamically, PAH is defined by precapillary pulmonary hypertension with an elevated mean pulmonary arterial pressure (mPAP) > 20 mmHg, a pulmonary arterial wedge pressure \leq 15 mmHg, and an elevated pulmonary vascular resistance (PVR) > 2 Wood units in right heart catheterization (RHC) at rest [5, 6].

Despite the progress in therapy, which dramatically improved clinical outcome, complete disease control is often not achieved and long-term survival of PAH patients remains low [7, 8]. PAH treatment decisions are based on individualized risk assessment at diagnosis and follow-up [5, 9]. Still, optimal PAH risk management has not been achieved yet, and available PAH risk stratification tools are highly heterogeneous, differ in the choice, number and weighting of used parameters, and the risk class definition [1, 8–13]. It has been recognized that the current definition of PAH includes heterogeneous patient phenotypes which differ in treatment responses and clinical outcomes [14, 15]. Following this observation, typical and atypical forms of PAH have been described, whereas the latter term refers to a late-onset PAH in patients with multiple comorbidities as compared to typical PAH, which is found in younger patients with few comorbidities [14]. Such phenotyping of PAH is essential to further improve individualized risk stratification and treatment decisions, which is addressed by numerous recent reports [5]. Current research on PAH risk assessment focuses primarily on novel risk parameters including imaging, functional, genetic, and proteomic features [16, 17]. Still, the implementation of such new risk markers in clinical practice is often hindered by a lack of availability, standardization, and increasing costs. By contrast, there are several more easily accessible features, such as age and sex, PAH etiology, serological parameters, pulmonary function tests, cardiac and pulmonary imaging, electrocardiography, exercise testing, and RHC measures which have been shown to predict mortality in PAH but are inconsistently used in currently available PAH risk scores [18–20]. Thus, we hypothesized that a machine learning approach employing e. g. Elastic Net survival modeling and clustering based on a broad set of easily accessible demographic and clinical parameters available at PAH diagnosis may facilitate risk stratification and identification of clinically relevant subsets of PAH patients.

Methods

Ethics

All participants gave written informed consent to participate. The study data were stored and analyzed in anonymized form. The study was approved by the ethics

committees of the Medical University of Innsbruck (approval numbers: AM2544, 239/4.12 and 273/5.7), the Johannes Kepler University of Linz (AN2017-0,009,369/4.15) and the Medical University of Vienna (EKV516/2011) and conducted in accordance with the Declaration of Helsinki and European data policy.

Study population and design

We herein present a retrospective multicenter observational two-cohort study. Data of 183 PAH patients were analyzed. The inclusion criterion was PAH (WHO etiology group I) defined by ESC/ERS (European Society of Cardiology/European Respiratory Society) guidelines and confirmed by RHC. The exclusion criterion was incompatibility with PAH diagnosis, e. g. post-capillary pulmonary hypertension or pulmonary hypertension group II, III, IV, or V. The participants were grouped in (1) the training cohort recruited at the Medical University of Innsbruck, Austria, (IBK, $N=100$) and the (2) test cohort recruited at the Elisabethinen Hospital Linz, Austria, and the Medical University of Vienna, Austria (LZ/W, $N=83$) (Fig. 1).

Study procedures and variables

Demographic and performance variables, blood biomarkers, heart echocardiography, and RHC parameters were determined during the standard PAH diagnosis procedure. Capillary blood gas analysis was performed by puncture of the hyper-perfused earlobe (induced by Finalgon® application [Sanofi-Aventis, Germany]). The glomerular filtration rate (GFR) was calculated by the MDRD GFR equation, renal insufficiency was defined as $GFR < 60$ ml/min/1.73 m². Anemia was defined by a hemoglobin concentration < 120 g/L for women and < 130 g/L for men. Impaired oxygenation was defined as peripheral hemoglobin oxygen saturation (SO₂) < 95%. The list of analyzed variables with the stratification scheme is presented in Supplementary Table S1.

Mortality risk-assessment tools

The mortality risk at diagnosis was assessed by abbreviated versions of the ESC/ERS risk assessment tool: the three- and four parameter French Pulmonary Hypertension Registry scores (FPHR 3p and 4p), the Prospective Registry of Newly Initiated Therapies for Pulmonary Hypertension (COMPERA) tool, the Swedish PAH register (SPAHR) model, and the modified Risk Assessment Score of PAH (mRASP) [8, 13, 21, 22].

Statistical analysis

Statistical analysis was performed with R version 4.2.3 (R Foundation for Statistical Computing).

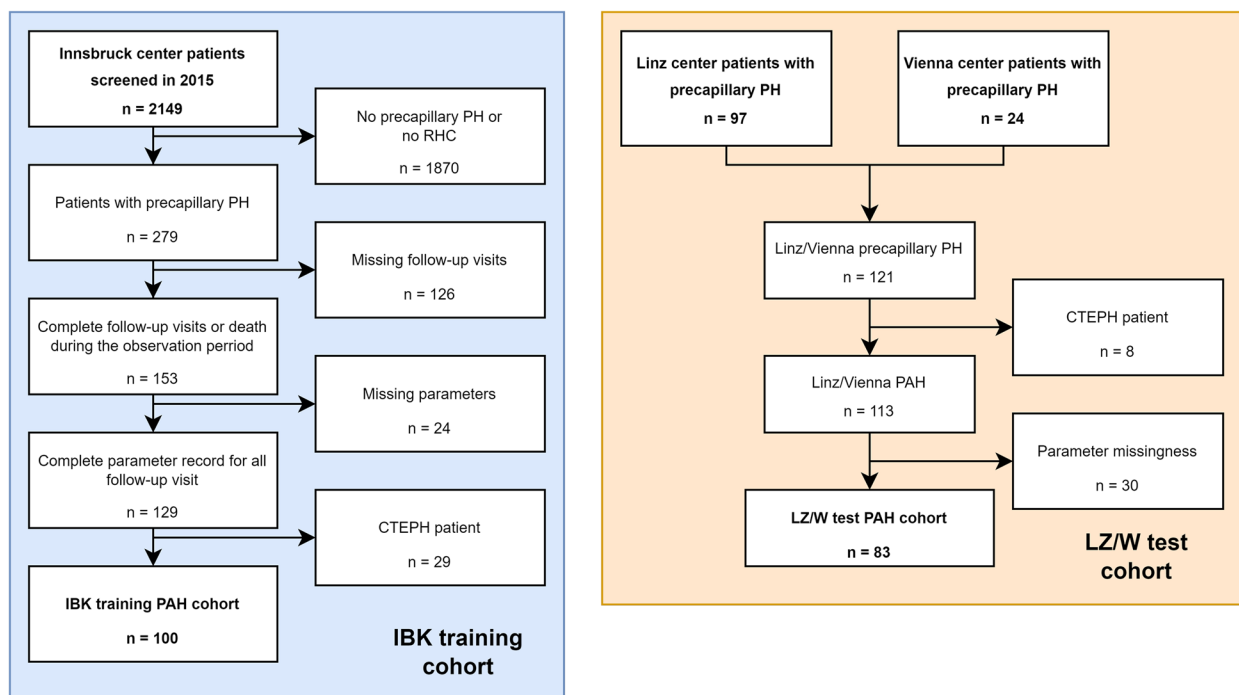


Fig. 1 Flow diagram of the study analysis inclusion process. PH: pulmonary hypertension; RHC: right heart catheterization; CTEPH: chronic thromboembolic pulmonary hypertension

In descriptive statistic, numeric variables are presented as medians with interquartile ranges (IQR). Categorical variables are presented as percentages and counts for each category. Prior to modeling and clustering, numeric variables were normalized and median centered. Both first and second order terms for numeric variables were included in models to account for non-linear relationship with survival.

Statistical significance was assessed by Mann–Whitney, χ^2 , log-rank and Spearman’s test for numeric variables, categorical parameters, survival and correlation, respectively. *P* values were adjusted for multiple testing within each analysis and cohort with the Benjamini–Hochberg method [23]. Association of single factors (Supplementary Table S1) or established risk assessment tools (mRASP, SPAHR, COMPERA, FPHR) with overall survival was investigated by univariable Cox proportional hazard modeling [24]. For multi-parameter survival modeling, the Elastic Net model [25, 26] was trained in the IBK cohort with the optimal lambda parameter obtained in 200-repetition tenfold cross-validation. Subsequently, the Elastic Net model linear predictor scores were calculated for the training IBK and test LZ/W cohort and their association with overall survival was assessed by univariable Cox modeling [24, 27, 28]. Performance of survival models was measured by concordance index (C) [29], integrated Brier score

(IBS) [30] and R2 explained variance statistics. Non-zero model coefficients in the Elastic Net model (hazard ratio \neq 1) were deemed the ‘Elastic Net signature’. Clustering of the IBK cohort participants in respect to the Elastic Net signature variables was done with PAM (partition around medoids) algorithm and cosine distance measure [31, 32]. The optimal cluster number was determined by the bend of the within-cluster sum of squares curve, the peak silhouette statistic [33], and the largest cluster assignment accuracy in tenfold cross-validation [34]. Cluster assignment in the LZ/W collective was accomplished by the inverse distance-weighted 7-nearest neighbor classifier [35]. Details of statistical analysis are provided in [Supplementary Methods](#). The analysis pipeline is available at <https://github.com/PiotrTymoszuk/PAH-biomarker>.

Results

Characteristics of the study cohorts

The PAH study cohorts (total: *N* = 183, IBK: *N* = 100, LZ/W: *N* = 83, Fig. 1) encompassed subjects with idiopathic (*N* = 172, 94%), hereditary (*N* = 3, 1.6%) and connective tissue disease associated PAH (*N* = 8, 4.4%). The latter group consisted of patients with systemic sclerosis, systemic lupus erythematosus, Sjörgeren syndrome and mixed connective tissue disease. The median age was 66 years [IQR: 53 – 71] in the IBK and 70 years [IQR: 54

– 74] in the LZ/W collective. The cohorts included predominantly females (IBK: 64% and LZ/W: 66%) (Table 1). We found no significant differences in observation time ($p=0.30$, Mann–Whitney test), overall survival ($p=0.51$, log-rank test) and overall mortality rate ($p=0.35$, χ^2 test) between the IBK and LZ/W cohorts. Accordingly, the median overall survival defined as follow-up time between the diagnosis and the last visit or death was 70 months [IQR: 46 – 110] in the IBK and 63 months [IQR: 32 – 110] in the LZ/W cohort, and mortality during follow-up was 33% in the IBK and 24% in the LZ/W cohort (Table 1, Supplementary Tables S1 – S2).

Univariable survival modeling

Initially, we applied univariable Cox modeling to search for survival-associated factors among 19 demographic, performance, biochemical and cardiopulmonary parameters measured at diagnosis (Supplementary Table S1). By this approach, N-terminal pro-brain natriuretic peptide (NT-pro-BNP), pulmonary vascular resistance (PVR) and six-minute walking distance (SMWD) were identified as significant survival-associated factors in both the IBK and LZ/W cohorts. Additionally, age, mean corpuscular volume, right atrial area (RAA), renal insufficiency were linked to a worse prognosis and cardiac index (CI) was associated with better survival in the IBK collective. In

the LZ/W cohort, male sex and III/IV WHO functional class were identified as unfavorable prognostic factors (Supplementary Figure S2, Supplementary Table S3).

Development of a multi-parameter PAH risk signature with Elastic Net modeling

Multi-parameter modeling of overall survival was performed with Elastic Net Cox regression [25, 26]. In the training IBK cohort, 7 out of 19 candidate independent variables, namely age, CI, NT-pro-BNP, PVR, RAA, RDW (red blood cell distribution width) and SMWD, contributed to risk prediction as reflected by hazard ratio $\neq 1$ and constituted the ‘Elastic Net signature’. In particular, age at diagnosis (linear: hazard ratio [HR]=1.16, quadratic term: HR=1.17) and NT-pro-BNP (linear term: HR=1.19) were the strongest predictors of an unfavorable clinical course, whereas CI (linear: HR=0.843, quadratic term: 0.95) and SMWD (linear: 0.902, quadratic term: 0.905) were linked to better overall survival (Fig. 2A).

The Elastic Net signature displayed good survival prediction accuracy [29, 30] in the training IBK ($C=0.82$, $R^2=0.65$, $IBS=0.098$) and the test LZ/W collective ($C=0.77$, $R^2=0.52$, $IBS=0.11$) (Supplementary Table S4). Accordingly, the linear predictor tertiles of the Elastic Net signature were associated with a low, intermediate

Table 1 Characteristics of the Innsbruck (IBK) and Linz/Vienna (LZ/W) study cohorts. Numeric variables are presented as medians with interquartile ranges (IQR) and ranges. Categorical variables are presented as percentages and counts within the complete observation set

Variable ^a	IBK	LZ/W	Significance ^b	Effect size ^b
Participants, n	100	83		
Age, years	66 [IQR: 53–71] 19–84	70 [IQR: 54–74] 23–82	ns ($p=0.36$)	$r=0.093$
Sex	female: 64% (64) male: 36% (36)	female: 66% (55) male: 34% (28)	ns ($p=0.94$)	$V=0.024$
Anemia	19% (19)	17% (14)	ns ($p=0.94$)	$V=0.028$
Renal insufficiency	35% (35)	18% (15)	$p=0.043$	$V=0.19$
Pericardial effusion	16% (16)	3.6% (3)	$p=0.04$	$V=0.2$
WHO class	I/II: 39% (39) III/IV: 61% (61)	I/II: 53% (44) III/IV: 47% (39)	ns ($p=0.17$)	$V=0.14$
SMWD, m	320 [IQR: 200–400] 50–610	350 [IQR: 270–440] 50–620	ns ($p=0.11$)	$r=0.15$
mPAP, mmHg	40 [IQR: 30–50] 26–120	39 [IQR: 31–49] 18–67	ns ($p=0.91$)	$r=0.024$
PVR, Wood	10 [IQR: 6.7–17] 3.3–43	5 [IQR: 3.5–7.8] 1.4–20	$p<0.001$	$r=0.54$
5-year mortality	21% (21)	13% (11)	ns ($p=0.38$)	$V=0.1$
OS, months	70 [IQR: 46–110] 2–230	63 [IQR: 32–110] 11–170	ns ($p=0.51$)	

^a WHO class: WHO functional class; SMWD: six-minute walking distance; mPAP: mean pulmonary arterial pressure; PVR: pulmonary vascular resistance; OS: overall survival

^b Numeric variables: Mann–Whitney U test with r effect size statistic; categorical variables: χ^2 test with Cramer V effect size statistic; survival: log-rank test

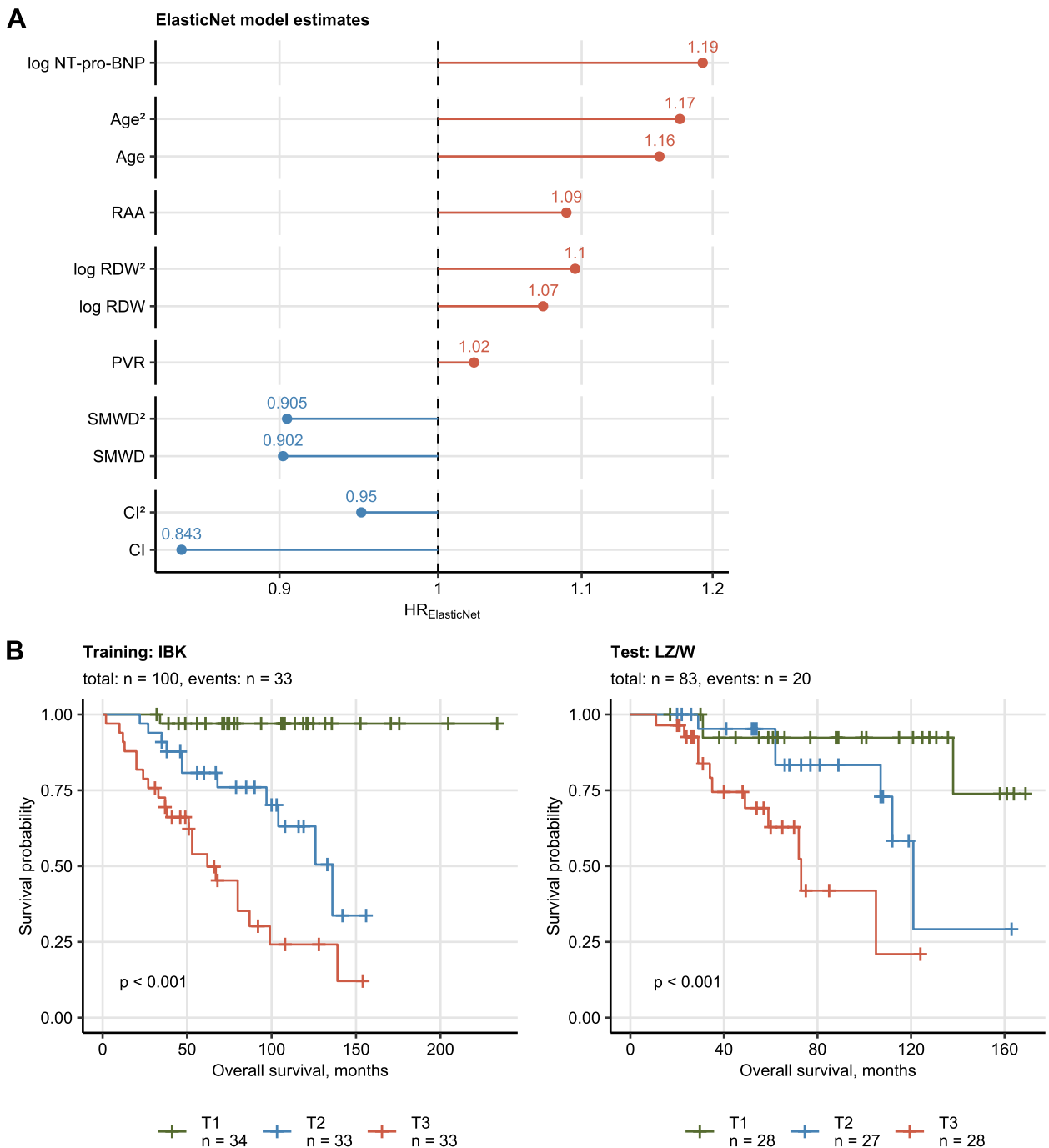


Fig. 2 Multi-parameter modeling of PAH survival with Elastic Net Cox regression. The Elastic Net multi-parameter Cox regression model with the set of 19 (Supplementary Table S1) independent variables and overall survival as a response was developed in the training Innsbruck cohort. Numeric independent variables were median centered and their first and second order terms included in the model. Numbers of complete observations and mortality is indicated in **B**. **A** Non-zero Elastic Net model coefficients (Elastic Net signature) represented as hazard ratios (HR). Plot points are labeled with their HR values. **B** Association of overall survival with the Elastic Net model linear prediction score in the training IBK and test Linz/Vienna (LZ/W) cohort was assessed by Kaplan–Meier analysis. Significance of the survival differences in the study participants stratified by the linear predictor score tertiles (T1: 0–33, T2: 34–66, T3: 66–100 percentile) was determined by log-rank test adjusted for multiple testing with Benjamini–Hochberg method. *P* values are shown in the plots, numbers of complete observations and mortality are indicated in the plot captions

and high risk of overall mortality in the Kaplan–Meier analysis in both cohorts indicating proper model calibration (Fig. 2B).

Prediction of overall survival by the Elastic Net signature and established PAH risk assessment tools

The Elastic Net signature values correlated significantly with risk class assignment by mRASP, COMPERA and SPAHR, and a number of risk factors in the FPHR models. However, the strength of correlation of the Elastic Net signature with other tools (Spearman’s ρ , IBK: 0.57 – 0.7, LZ/W: 0.7 – 0.79) tended to be lower than the correlation of the established risk tools with each other (IBK: 0.61 – 0.9, LZ/W: 0.68 – 0.9) (Supplementary Figure S2). This suggests that the Elastic Net signature may provide better risk estimates for PAH individuals with inadequate survival prediction by other tools.

To test for that, we compared predictive performance of the newly developed Elastic Net signature with the established risk assessment tools and an ensemble of the established risk scales developed with the Ridge Cox technique [25, 26] (Supplementary Figure S4). In such comparison, the Elastic Net signature displayed a better

prediction accuracy gauged by C-index, IBS and R2 than the best performing comparator risk scales. Furthermore, although the performance of the ensemble model and the Elastic Net signature in the IBK cohort was comparable, the ensemble of established risk scales performed substantially worse in the LZ/W collective (Fig. 3, Supplementary Table S4).

Identification of PAH risk phenotypes by clustering

Next, we investigated if the Elastic Net signature variables may be applied for definition of clinically relevant subsets of PAH. To this end, we assigned participants of the training IBK cohort to two clusters defined by the PAM algorithm with cosine distance [31, 32] in respect to age, CI, NT-pro-BNP, PVR, RAA, RDW and SMWD. The clustering algorithm of choice displayed the superior reproducibility in tenfold cross-validation [34] (accuracy = 0.97) and high explanatory value (‘explained’ variance fraction = 0.56) as compared with several other procedures such as hierarchical or k-means algorithm. The two cluster solution was also optimal in terms of explanatory value and reproducibility as tested for PAM/cosine distance clustering structures with varying cluster

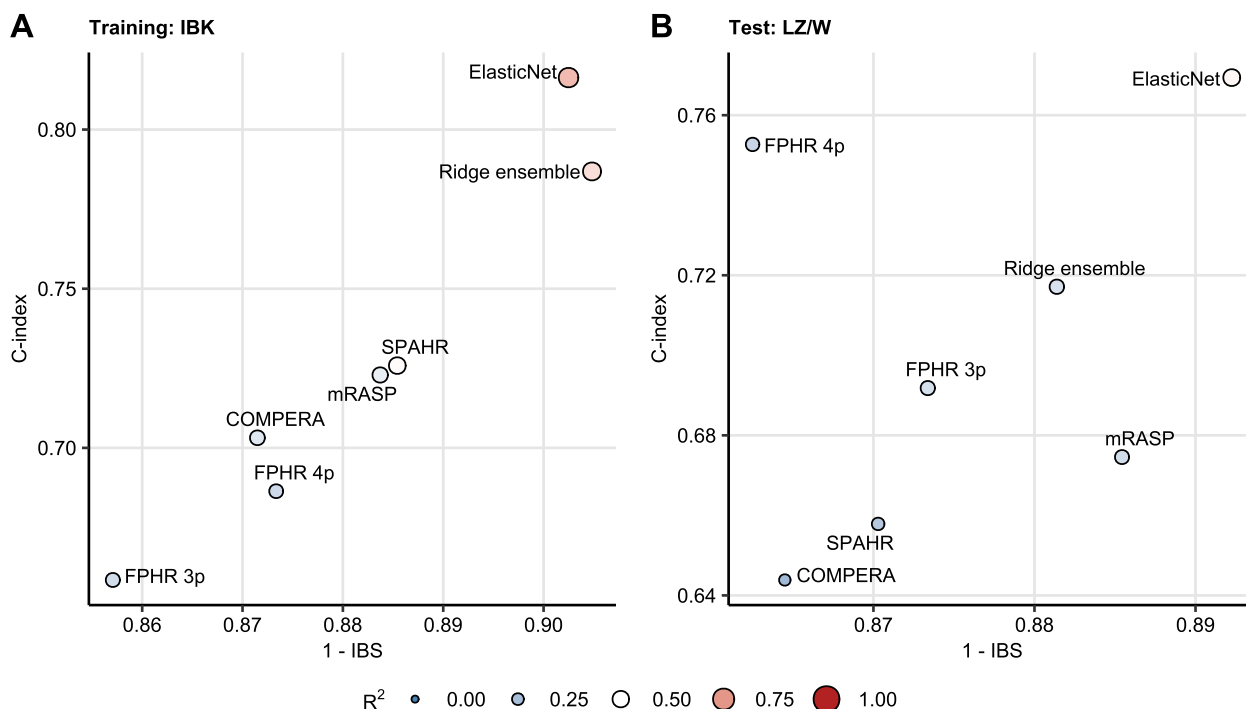


Fig. 3 Performance of PAH risk assessment tools. The Elastic Net signature was developed as presented in Fig. 2. The ensemble of the established risk assessment tools (FPHR 3p: French Pulmonary Hypertension Registry 3 parameter score, FPHR 4p: French Pulmonary Hypertension Registry 4 parameter score, COMPERA: Comparative, Prospective Registry of Newly Initiated Therapies for Pulmonary Hypertension score, mRASP: modified Risk Assessment Score of PAH) was developed by Ridge Cox regression as presented in Supplementary Figure S3. Predictive performance of the Elastic Net signature, 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, point size and color codes for R²

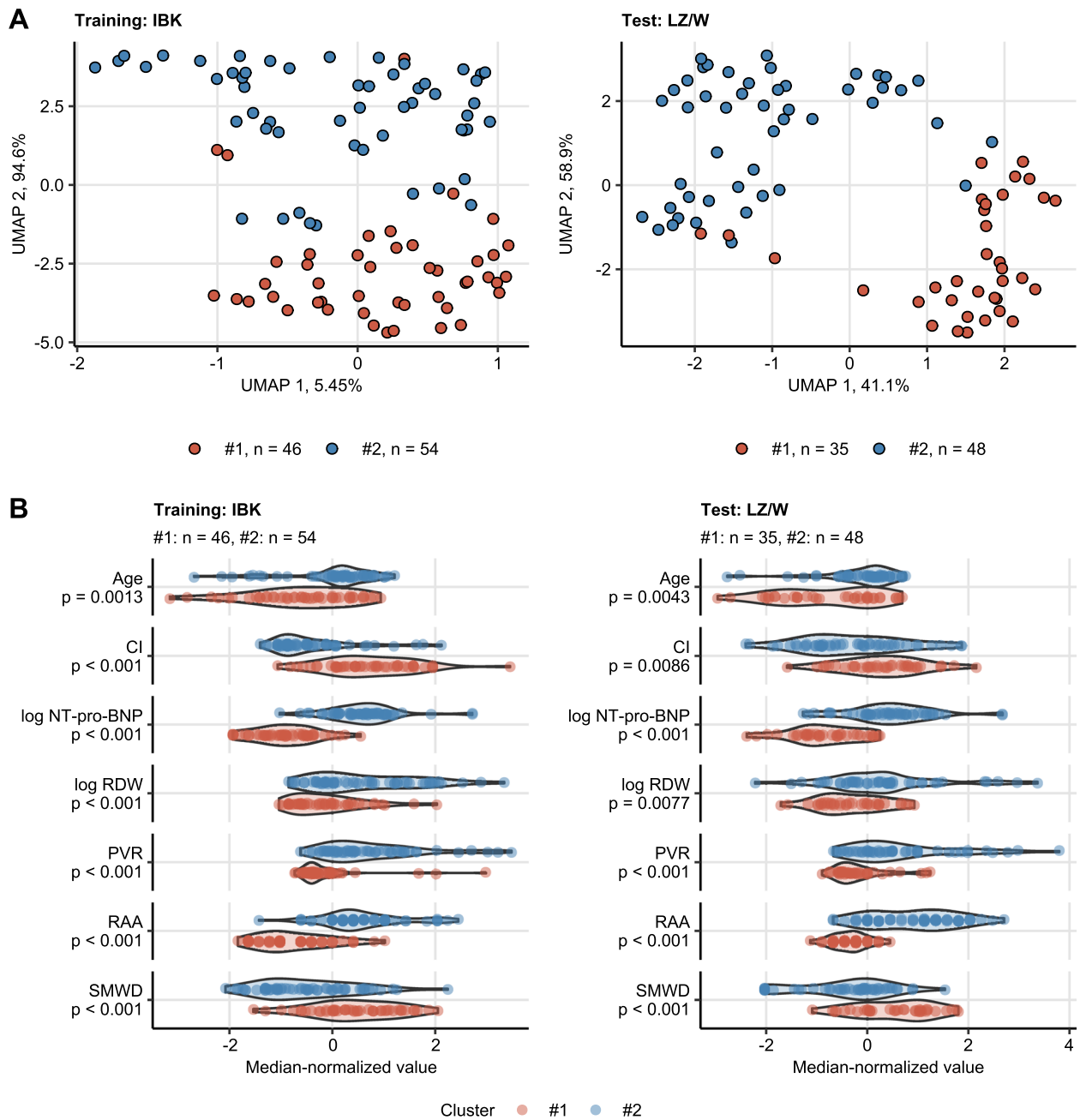


Fig. 4 Clustering of the study participants. Clustering of the training Innsbruck (IBK) cohort participants in respect to the survival-associated factors identified by Elastic Net modeling (Fig. 2) was investigated by PAM (partition around medoids) algorithm with cosine distance. Numeric clustering features were median centered prior to the clustering. Cluster assignment in the training Linz/Vienna cohort (LZ/W) was done by an inverse distance weighted 7-nearest neighbor classifier. Numbers of individuals assigned to the PAH clusters are presented in the plot captions or legends. **A** PAH cluster assignment overlaid on the 2-dimensional cosine-distance UMAP (Uniform Manifold Approximation and Projection for Dimension Reduction) layout plots. Percentages of variance associated with the components are indicated in the plot axes. **B** Differences in the clustering features between the PAH clusters were assessed by Mann–Whitney test corrected for multiple testing with Benjamini–Hochberg method. Normalized, median-centered values of the clustering factors are shown in violin plots. Points represent single observations. *P* values are indicated in the Y axes. CI: cardiac index; NT-pro-BNP: N terminal pro brain natriuretic peptide; RDW: red blood cell distribution width; PVR: pulmonary vascular resistance; RAA: right atrial area; SMWD: six minute walking distance

numbers. NT-pro-BNP, CI and RAA were found to be the most influential clustering factors (Supplementary Figures S4 – S5). The cluster assignment in the test LZ/W cohort was accomplished by a 7-nearest neighbors classifier (Fig. 4A). In both study collectives, the smaller PAH cluster #1 (IBK: 46%, LZ/W: 42% participants) encompassed significantly younger participants with better cardiac function reflected by a higher CI, lower RAA and lower blood NT-pro-BNP levels than cluster #2 participants. Furthermore, cluster #1 participants had lower PVR, lower RDW and superior SMWD in comparison with PAH cluster #2 (Fig. 4B). Additionally, mean pulmonary (mPAP) and right atrial (mRAP) pressure were significantly higher in PAH cluster #2 than in cluster #1 participants in both the IBK and LZ/W cohorts (Supplementary Figure S6, Supplementary Tables S4 – S5).

Finally, in line with the poorer physical performance and higher age, PAH cluster #2 demonstrated significantly worse risk profile composition with a higher number of risk factors in the FPHR model and more patients assigned to intermediate and high-risk strata by the mRASP, COMPERA and SPAHR tools. Consequently, overall survival in PAH cluster #2 was significantly shorter than in PAH cluster #1 (Fig. 5, Supplementary Figure S6, Supplementary Tables S5 – S6).

Discussion

In PAH, standardized risk scores are used to evaluate the risk of mortality at initial presentation of the disease, and further to guide and adjust therapy upon a follow-up risk re-assessment and definition of treatment goals (e. g. low-risk status) [5]. Consequently, accurate risk assessment is key to the successful PAH management and various risk assessment strategies have been proposed [18–20]. While some approaches favor simple scores eligible for repetitive clinical evaluation, others employ numerous and sometimes not easily available parameter sets for high prediction accuracy [9]. Interestingly, currently available risk scores are mainly based on expert opinion and traditional hypothesis-driven statistical tools such as logistic regression [9].

In this proof of concept study, we demonstrate the usefulness of supervised multi-parameter Elastic Net survival modeling [25, 26] and unsupervised PAM clustering [32] for search of novel biomarker combinations and

subsets of PAH patients, which may improve PAH survival prediction. As PAH is a rare disease, patients from three Austrian PAH centers were included in our study. This design enabled us for an external validation of the modeling and clustering results. Our findings corroborate previous studies, reporting high accuracy of risk prediction and informative phenotyping of PAH patients by machine learning and clustering algorithms [15, 36, 37].

By Elastic Net regression, we generated a highly accurate and reproducible model, which outcompeted single demographic, biochemical and functional factors, such as the widely used COMPERA, SPAHR, mRASP and FPHR models [5, 8, 13, 21, 22] as well as the ensemble model combining those popular risk scales [26]. The Elastic Net signature comprises well described risk factors of overall PAH mortality such as NT-pro-BNP, age, but also impaired functional parameters, including a reduced six-minute walking distance, as well as impaired hemodynamics, such as poor cardiac output and high pulmonary vascular resistance. However, those parameters are not consistently included in established risk scores [9, 15]. Most prominently, advanced age, which was found to be a strong unmodifiable mortality predictor in the analyzed collectives both in the uni- and multi-variable setting, is not part of most risk assessment schemes including COMPERA, SPAHR, mRASP and FPHR. Another easily accessible parameter, RDW, was associated with poorer survival in Elastic Net modeling, which underscores the clinically relevant link between iron turnover, inflammation, oxidative stress, erythropoiesis and PAH progression [38–41]. Of note, although the correlation of RDW and PAH mortality has been reported before [42], this risk factor is not routinely determined by automated blood count systems and has not been prospectively validated yet.

Our clustering scheme utilizing Elastic Net signature variables reproduces previously published phenotypes of typical and atypical PAH and demonstrates significant differences in the long-term follow-up of these patients [14]. In detail, PAH cluster #2 encompassed mainly the elderly and, presumably, comorbid individuals with a shorter overall survival as compared to typical PAH patients preferentially assigned to cluster #1.

Our study has some limitations. First, as PAH is a rare condition, our study followed a multi-center retrospective

(See figure on next page.)

Fig. 5 Risk assessment and survival differences in the PAH clusters. Risk assessment strata distribution and overall survival was compared between the study participant clusters with and log-rank test, respectively. *P* values were adjusted for multiple testing with Benjamini–Hochberg method. Numbers of individuals assigned to the clusters are presented in the Y axes or in the plot legends. **A – E** Risk assessment strata frequencies in the PAH clusters presented in stack plots. All differences were significant with $p < 0.001$. **F** Differences in overall survival in the PAH clusters visualized in Kaplan–Meier plots. *P* values are indicated in the plots. Numbers of complete observations and deaths are shown in the plot captions. 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; int: intermediate

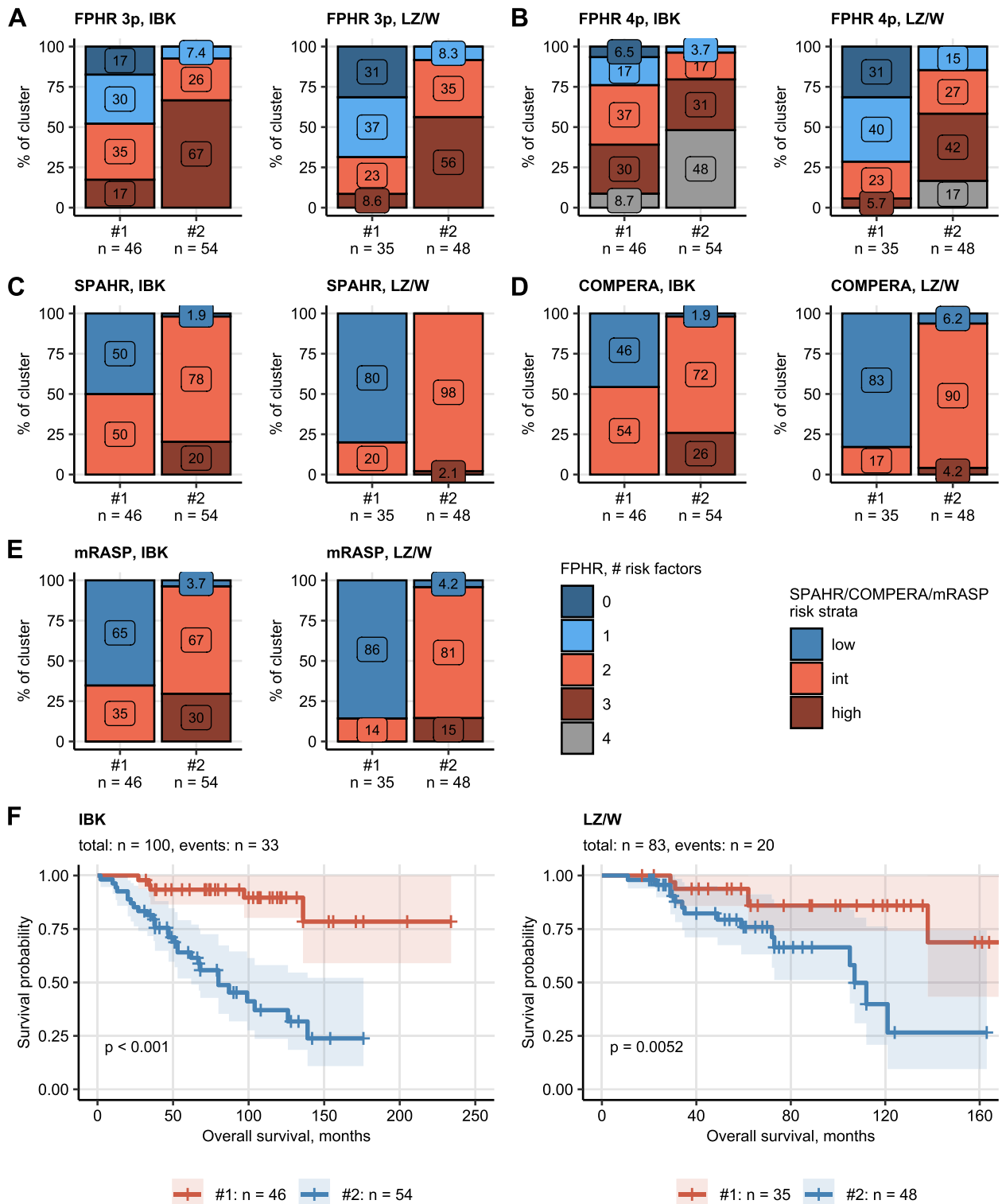


Fig. 5 (See legend on previous page.)

design, which yielded only moderate size cohorts. Hence, the Elastic Net signature and PAH clustering scheme developed and tested in the IBK and LZ/W cohort need to be validated in larger prospective studies. A larger cohort size

and algorithms robustly handling interactions between explanatory variables such as neuronal networks or tree models [36, 37] may improve the survival model accuracy even further. Second, although we screened a relatively large

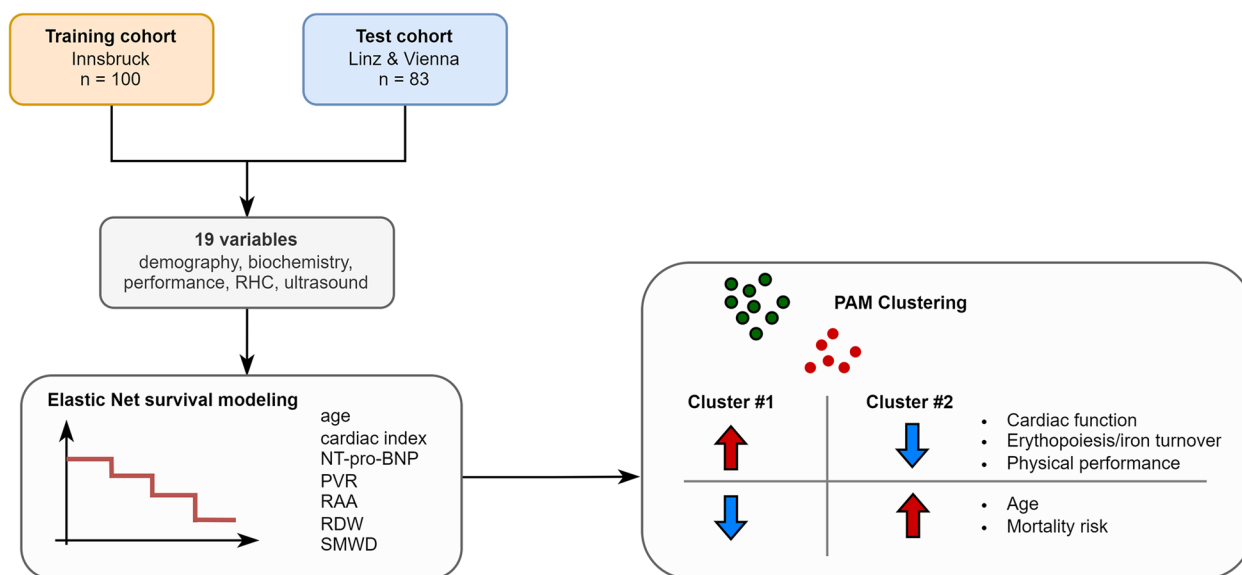


Fig. 6 Summary of the analysis results. RHC: right heart catheterization; Cl: cardiac index; NT-pro-BNP: N terminal pro brain natriuretic peptide; RDW: red blood cell distribution width; PVR: pulmonary vascular resistance; RAA: right atrial area; SMWD: six minute walking distance

set of 19 explanatory factors compared with previous PAH risk analyses, we were not able to investigate some biomarkers of potential clinical relevance. For instance, diffusion capacity of carbon monoxide [15], was only available in the IBK cohort, and thus could not be included in the externally validated model. Additionally, we were not able to investigate new biomarkers, such as proteome profiling, as these parameters were not available for the presented cohorts [16, 17]. Although implementation of such new parameters may improve survival prediction or identify subsets of PAH patients at higher risk, their usefulness is limited by costs, need for worldwide standardization and scarce medical resources. Thus, we herein focused on broadly available candidate risk factors, rather than entirely new biomarkers. Still, our Elastic Net signature may be easily expanded and adapted to additional explanatory variables, thus providing a fast-track analysis tool for potential new biomarker sets in comparison to currently available risk models. Finally, our analysis included individuals at PAH diagnosis. Dynamics of the PAH clusters and modification of the risk predicted by Elastic Net modeling at follow-up and during treatment needs to be addressed by future research.

Conclusions

We herein provide proof of principle, that supervised and unsupervised learning algorithms may improve risk assessment, and identify clinically relevant subsets of patients, hence contributing to a better understanding of biologically distinct PAH phenotypes (Fig. 6). This may pave the way to effective individualized risk management and treatment in PAH.

Abbreviations

PAH	Pulmonary arterial hypertension
IBK	Innsbruck cohort
LZ/W	Linz/Vienna cohort
RHC	Right heart catheterization
mPAP	Mean pulmonary arterial pressure
PVR	Pulmonary vascular resistance
ESC/ERS	European Society of Cardiology/European Respiratory Society
GFR	Glomerular filtration rate
SO ₂	Peripheral hemoglobin oxygen saturation
FPHR	French Pulmonary Hypertension Registry
COMPERA	Prospective Registry of Newly Initiated Therapies for Pulmonary Hypertension
SPAHR	The Swedish PAH register
mRASP	Modified Risk Assessment Score of PAH
IQR	Interquartile range
C	C-index
IBS	Integrated Brier score
PAM	Partition around medoids
NT-pro-BNP	N-terminal pro-brain natriuretic peptide
SMWD	Six-minute walking distance
RAA	Right atrial area
CI	Cardiac index
RDW	Red blood cell distribution width

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12890-023-02427-2>.

Additional file 1.

Acknowledgements

We are grateful for the continued dedication of our staff to the care and treatment of pulmonary hypertension patients.

Authors' contributions

TS, PT, and JLR designed the study. TS, RSM, ES, SPS, IT, IL, and JLR collected data. TS and PT performed the statistical analysis. TS, RSM, GW, PT, and JLR

interpreted data. TS, PT, and JLR wrote the manuscript. The author(s) read and approved the final manuscript.

Authors' information

The final version of the manuscript was approved by all authors and we herewith verify that this manuscript has not been previously published. The manuscript is not under consideration or will be submitted elsewhere.

Funding

The work was funded by a grant from the Austrian National Bank Fund (Project 17271, JLR). The funding body had no influence on the design of the study and collection, analysis, and interpretation of data and writing of the manuscript.

Availability of data and materials

The study source data will be made available upon request. The source code of the analysis pipeline is available at <https://github.com/PiotrTymoszuk/PAH-biomarker>.

Declarations

Ethics approval and consent to participate

The study was conducted in accordance with the Declaration of Helsinki. All participants gave written informed consent to participate. The study was approved by the ethics committees of the Medical University of Innsbruck (approval numbers: AM2544, 239/4.12 and 273/5.7), the Johannes Kepler University of Linz (AN2017-0009369/4.15) and the Medical University of Vienna (EKV516/2011).

Consent for publication

Does not apply.

Competing interests

PT owns a data science enterprise Data Analytics As a Service Tirol. Other authors declare no conflict of interest connected with this study.

Author details

¹Department of Internal Medicine II, Medical University of Innsbruck, Anichstraße 35, 6020 Innsbruck, Austria. ²Data Analytics As a Service Tirol, Daas.Tirol, Innsbruck, Austria. ³Department of Cardiology, Elisabethinenkrankenhaus, Linz, Austria. ⁴Department of Cardiology, Medical University of Vienna, Vienna, Austria.

Received: 11 October 2022 Accepted: 6 April 2023

Published online: 25 April 2023

References

- Sitbon O, Benza RL, Badesch DB, Barst RJ, Elliott CG, Gressin V, et al. Validation of two predictive models for survival in pulmonary arterial hypertension. *Eur Respir J*. 2015;46:152–64.
- D'Alonzo GE, Barst RJ, Ayres SM, Bergofsky EH, Brundage BH, Detre KM, et al. Survival in patients with primary pulmonary hypertension. Results from a national prospective registry. *Ann Intern Med*. 1991;115:343–9.
- Hoeper MM, Humbert M, Souza R, Idrees M, Kawut SM, Sliwa-Hahnle K, et al. A global view of pulmonary hypertension. *Lancet Respir Med*. 2016;4:306–22.
- Farber HW, Miller DP, Poms AD, Badesch DB, Frost AE, Muros-Le Rouzic E, et al. Five-Year outcomes of patients enrolled in the REVEAL Registry. *Chest*. 2015;148:1043–54.
- Humbert M, Kovacs G, Hoeper MM, Badagliacca R, Berger RMF, Brida M, et al. 2022 ESC/ERS Guidelines for the diagnosis and treatment of pulmonary hypertension. *Eur Respir J*. 2023;61(1):2200879. <https://doi.org/10.1183/13993003.00879-2022>.
- Simonneau G, Montani D, Celermajer DS, Denton CP, Gatzoulis MA, Krowka M, et al. Haemodynamic definitions and updated clinical classification of pulmonary hypertension. *Eur Respir J*. 2019;53:1801913.
- Humbert M, Sitbon O, Yaïci A, Montani D, O'Callaghan DS, Jaïs X, et al. Survival in incident and prevalent cohorts of patients with pulmonary arterial hypertension. *Eur Respir J*. 2010;36:549–55.
- Hoeper MM, Kramer T, Pan Z, Eichstaedt CA, Spiesshoefer J, Benjamin N, Mortality in pulmonary arterial hypertension: prediction by the, et al. European pulmonary hypertension guidelines risk stratification model. *Eur Respir J*. 2015;2017:50.
- Humbert M, Farber HW, Ghofrani HA, Benza RL, Busse D, Meier C, et al. Risk assessment in pulmonary arterial hypertension and chronic thromboembolic pulmonary hypertension. *Eur Respir J*. 2019;53:1802004.
- 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. *Chest*. 2019;156:323–37.
- Benza RL, Gomberg-Maitland M, Miller DP, Frost A, Frantz RP, Foreman AJ, et al. The REVEAL Registry risk score calculator in patients newly diagnosed with pulmonary arterial hypertension. *Chest*. 2012;141:354–62.
- Thenappan T, Shah SJ, Rich S, Tian L, Archer SL, Gomberg-Maitland M. Survival in pulmonary arterial hypertension: a reappraisal of the NIH risk stratification equation. *Eur Respir J*. 2010;35:1079–87.
- Boucly A, Weatherald J, Savale L, Jaïs X, Cottin V, Prevot G, et al. Risk assessment, prognosis and guideline implementation in pulmonary arterial hypertension. *Eur Respir J*. 2017;50:1700889.
- Hoeper MM, Apitz C, Grünig E, Halank M, Ewert R, Kaemmerer H, et al. Targeted therapy of pulmonary arterial hypertension: Updated recommendations from the Cologne Consensus Conference 2018. *Int J Cardiol*. 2018;272S:37–45.
- Hoeper MM, Pausch C, Grünig E, Klose H, Staehler G, Huscher D, et al. Idiopathic pulmonary arterial hypertension phenotypes determined by cluster analysis from the COMPERA registry. *J Hear Lung Transplant*. 2020;39:1435–44.
- Rhodes CJ, Wharton J, Swietlik EM, Harbaum L, Gireb B, Coghlan JG, et al. Using the Plasma Proteome for Risk Stratifying Patients with Pulmonary Arterial Hypertension. *Am J Respir Crit Care Med*. 2022. <https://doi.org/10.1164/RCCM.202105-1118OC>.
- Amsallem M, Sweatt AJ, Ataam JA, Guihaire J, Lecerf F, Lambert M, et al. Targeted proteomics of right heart adaptation to pulmonary arterial hypertension. *Eur Respir J*. 2021;57:2002428.
- Geenen LW, Baggen VJM, Koudstaal T, Boomars KA, Eindhoven JA, Boersma E, et al. The prognostic value of various biomarkers in adults with pulmonary hypertension; a multi-biomarker approach. *Am Heart J*. 2019;208:91–9.
- Foris V, Kovacs G, Tscherner M, Olschewski A, Olschewski H. Biomarkers in pulmonary hypertension: what do we know? *Chest*. 2013;144:274–83.
- Galiè N, Channick RN, Frantz RP, Grünig E, Jing ZC, Moiseeva O, et al. Risk stratification and medical therapy of pulmonary arterial hypertension. *Eur Respir J*. 2019;53:1801889.
- Xiong W, Zhao Y, Xu M, Pudasaini B, Guo X, Liu J. A modified risk score in one-year survival rate assessment of group 1 pulmonary arterial hypertension. *BMC Pulm Med*. 2018;18:161.
- Kyllhammar D, Kjellström B, Hjalmarsson C, Jansson K, Nisell M, Söderberg S, et al. A comprehensive risk stratification at early follow-up determines prognosis in pulmonary arterial hypertension. *Eur Heart J*. 2018;39:4175–81.
- Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J R Stat Soc Ser B*. 1995;57:289–300.
- Therneau TM, Grambsch PM. *Modeling Survival Data: Extending the Cox Model*. 1st ed. New York: Springer Verlag; 2000.
- Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*. 2010;33:1–22.
- Simon N, Friedman J, Hastie T, Tibshirani R. Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *J Stat Softw*. 2011;39:1–13.
- Royston P, Altman DG. External validation of a Cox prognostic model: Principles and methods. *BMC Med Res Methodol*. 2013;13:33.
- Simon RM, Subramanian J, Li MC, Menezes S. Using cross-validation to evaluate predictive accuracy of survival risk classifiers based on high-dimensional data. *Brief Bioinform*. 2011;12:203–14.
- Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med*. 1996;15:361–87.

30. Graf E, Schmoor C, Sauerbrei W, Schumacher M. Assessment and comparison of prognostic classification schemes for survival data. *Stat Med*. 1999;18:2529–45.
31. Drost H-G. Philentropy: Information Theory and Distance Quantification with R. *J Open Source Softw*. 2018;3:765.
32. Schubert E, Rousseeuw PJ. Faster k-Medoids Clustering: Improving the PAM, CLARA, and CLARANS Algorithms. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Springer; 2019. p. 171–87.
33. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math*. 1987;20 C:53–65.
34. Lange T, Roth V, Braun ML, Buhmann JM. Stability-based validation of clustering solutions. *Neural Comput*. 2004;16:1299–323.
35. Leng M, Wang J, Cheng J, Zhou H, Chen X. Adaptive semi-supervised clustering algorithm with label propagation. *J Softw Eng*. 2014;8:14–22.
36. Kanwar MK, Gomberg-Maitland M, Hoepfer M, Pausch C, Pittow D, Strange G, et al. Risk stratification in pulmonary arterial hypertension using Bayesian analysis. *Eur Respir J*. 2020;56:2000008.
37. Leha A, Hellenkamp K, Unsöld B, Mushemi-Blake S, Shah AM, Hasenfuß G, et al. A machine learning approach for the prediction of pulmonary hypertension. *PLoS One*. 2019;14:e0224453.
38. Rhodes CJ, Howard LS, Busbridge M, Ashby D, Kondili E, Gibbs JSR, et al. Iron deficiency and raised hepcidin in idiopathic pulmonary arterial hypertension: clinical prevalence, outcomes, and mechanistic insights. *J Am Coll Cardiol*. 2011;58:300–9.
39. Sonnweber T, Nairz M, Theurl I, Petzer V, Tymoszuk P, Haschka D, et al. The crucial impact of iron deficiency definition for the course of precapillary pulmonary hypertension. *PLoS One*. 2018;13:e0203396.
40. Salvagno GL, Sanchis-Gomar F, Picanza A, Lippi G. Red blood cell distribution width: a simple parameter with multiple clinical applications. *Crit Rev Clin Lab Sci*. 2015;52:86–105.
41. Thomas T, Pizzini A, Tancevski I, Löffler-Ragg J, Weiss G. Anaemia iron homeostasis and pulmonary hypertension: a review. *Abstract Inter Emerg Med*. 2020;15(4):573–85. <https://doi.org/10.1007/s11739-020-02288-1>.
42. 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. *Heart*. 2011;97:1054–60.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

