Risk stratification and phenotyping in pulmonary arterial hypertension with machine learning algorithms and cluster analysis – a long-term retrospective multicenter trial

Supplementary Material

# Supplementary Methods

## Data transformation, visualization, descriptive statistic

Data transformation, analysis and result visualization was accomplished by R version 4.0.5 with *tidyverse* environment1,2. Figures were generated with *cowplot* package,3 Supplementary Material file was built with *rmarkdown* environment (packages *knitr*, *rmarkdown*, *flextable* and *bookdown*).4

For univariable survival modeling and construction of candidate risk signatures, a set of categorical 19 demographic, biochemical, right-heart catheter, laboratory, ultrasound and lung function parameters recorded at PH diagnosis was used. To improve normality of some independent variables (NT-pro-BNP, RDW, TF-Sat, Ferritin) prior to survival modeling, log transformation was applied. For the list of modeling variables and their stratification scheme, see: **Supplementary Table 1**.

## Hypothesis testing, multiple comparisons

As the some of the analyzed numeric variables were non-normally distributed as checked by Shapiro-Wilk test, differences in median values of numeric variables between the study cohorts or participant clusters were investigated by Mann-Whitney test and r effect size statistic. Differences in frequency distribution of categorical variables between the study cohorts or participant clusters were assessed by test and Cramer’s V effect size statistic. Explorative data analysis and hypothesis testing was accomplished with *rstatix* package and in-house-developed tools (<https://github.com/PiotrTymoszuk/ExDA>). Differences in survival between the participant clusters or participants stratified by risk score tertiles were compared by Kaplan-Meier (KM) analysis, Mentel-Haenszel or log-rank test.5,6 KM analysis and visualization of its results were done with tools provided by *survival* and *survminer* packages.7 For each analysis and cohort, p values were corrected for multiple comparisons with Benjamini-Hochberg method.8

## Univariable Cox survival modeling

Association of independent categorical and numeric variables (**Supplementary Table 1**) with overall survival time was assessed by series of univariable Cox proportional hazard models constructed for the Innsbruck and Linz/Vienna cohort using *survival* package.6 Numeric variables were median-centered (function *scale(x, center = median(x))*). To account for non-linear associations of numeric independent variables, both 1st and 2nd order terms were included in the Cox models. Significance of the hazard ratio estimates was determined by Wald Z test. P values were corrected for multiple comparisons with Benjamini-Hochberg method.8 Proportional hazard assumption was checked with *cox.zph()* function (package *survival*).6 For the full modeling results, see: **Supplementary Table 3**.

## Multivariable Cox survival modeling with elastic net technique

Multi-parameter Cox modeling with the set of independent categorical and numeric variables (**Supplementary Table 1**) was accomplished by elastic net machine learning technique and *glmnet* package.9 Data pre-processing included median centering of numeric independent variables (function *scale(x, center = median(x))*) and conversion of categorical features to dummy numeric variables (function *model.matrix()*, base R). To account for non-linear associations of numeric independent variables, both 1st and 2nd order terms were included in the elastic net model development. The elastic net Cox proportional hazard model was trained in the Innsbruck cohort (function *glmnet()*, alpha = 0.5). The optimal lambda parameter ( = 0.166) for the training cohort model construction was found by 200-repetition 10-fold cross-validation (function *cv.glmnet()*) and corresponded to the minimum of cross-validation error. The values of non-zero elastic net model coefficients are presented in **Figure 2A**. Subsequently, the elastic net model linear predictor (LP) scores were calculated for the training IBK and test Linz/Vienna cohort and their association with overall survival was assessed by univariable Cox modeling. Concordance index (C-index) and (measure of explained variation) were calculated with *concordance()* (package *survival*) and *rsq()* (package *survMisc*) functions, respectively. Modeled survival in the training and the test cohort was compared with the actual overall survival by KM method. Differences in survival between study participants stratified by the LP score tertiles were assessed by log-rank test as described above.

## Clustering of the study participants

Clustering of the study participants in the Innsbruck training cohort in respect to the variables found associated with overall survival by the elastic net Cox modeling (**Figure 2A**; Age, SMWD, log RDW, CI, PVR, log NT-pro-BNP, RAA) was done with the PAM algorithm (partitioning around medoids, function *pam()*, package *cluster*)10 with the cosine distance between the study participants (function *distance()*, package *philentropy*).11 Prior to clustering, the numeric variables were median centered (function *scale(x, center = median(x))*). The PAM/cosine distance clustering procedure demonstrated the superior fraction of ‘explained’ clustering variance (ratio of total between-cluster to total sum-of-squares) and the optimal performance in 10-fold cross-validation measured by the fraction of correct cluster assignments12 as compared with hierarchical clustering, k-means and self-organizing map algorithms (**Supplementary Figure S2A**). The choice of cluster number was based on the bend of the within-cluster sum-of-squares curve (**Supplementary Figure 2B**). The importance of specific clustering features was determined by comparing the ‘explained’ clustering variances of the original clustering structure with the clustering objects with randomly re-shuffled clustering features.13 Assignment of the test Linz/Vienna cohort participants to the clusters was done with a k-nearest neighbor label propagation procedure (k = 5).13–15 In-house-developed wrappers for cluster object construction, cross-validation, importance testing and semi-supervised clustering are available as development packages (*clustTools*: <https://github.com/PiotrTymoszuk/clustTools> and *somKernels*: <https://github.com/PiotrTymoszuk/somKernels>).

Differences in study variables (**Supplementary Table 1**) between the participant clusters were determined by Mann-Whitney or test as described above (**Supplementary Tables S4 - S4**). Differences in overall survival between the clusters were compared with KM method and Mentel-Haenszel test as described above.

# Data and code availability

The study data set is available at serious request to the corresponding author. The analysis R code was deposited on GitHub (<https://github.com/PiotrTymoszuk/PAH-biomarker>).

# Supplementary Tables

**Supplementary** **Table S1:** Study variables.

| **Variable1** | **Description** | **Label2** | **Unit** | **Stratification** | **Used in risk modeling** |
| --- | --- | --- | --- | --- | --- |
| center | 1: Innsbruck, 2: Linz/Vienna | Cohort |  |  | no |
| ID | patient ID | ID |  |  | no |
| age\_fc | age at the diagnosis | Age | y |  | yes |
| SMWD | Six Minute Walk Distance | SMWD | m |  | yes |
| mPAP | Mean pulmonary arterial pressure | mPAP | mmHg |  | yes |
| Firstdiagnosisdate | date of first diagnosis, dd/mm/yyyy | Diagnosis date |  |  | no |
| event1 | 1-year mortality | 1-year mortality |  | no; yes | no |
| event3 | 3-year mortality | 3-year mortality |  | no; yes | no |
| event5 | 5-year mortality | 5-year mortality |  | no; yes | no |
| death\_study | overall mortality during the study period | Overall mortality |  |  | no |
| death\_study\_fct | overall mortality during the study period | Overall mortality |  | no; yes | no |
| Survival\_time\_from\_FD\_months | survival time from the diagnosis | OS | months |  | no |
| surv\_months | survival time from the diagnosis | OS | months |  | no |
| Date\_of\_death | death date | Death date |  |  | no |
| observation\_time\_yrs | observation time | Obs. Time | years |  | no |
| sex | sex | Sex |  | female; male | yes |
| PVR | Pulmonary vascular resistance | PVR | Wood |  | yes |
| PCWP | Pulmonary capillary wedge pressure | PCWP | mmHg |  | yes |
| anemia | anemia | Anemia |  | no; yes | yes |
| RDW\_log | RDW | log RDW | % |  | yes |
| renal\_ins | Renal insufficiency, GFR < 60% | Renal insufficiency |  | no; yes | yes |
| FT\_log | Ferritin | log FT | ng/ml |  | yes |
| TSAT\_log | Transferrin saturation | log TF-Sat | % |  | yes |
| MCV | MCV | MCV | fl |  | yes |
| NTproBNP\_log | NT-pro-BNP | log NT-pro-BNP | pg/ml |  | yes |
| percardial\_effusion | Percardial effusion | Percardial effusion |  | no; yes | yes |
| RA\_area | Right Atrial Area | RAA | cm2 |  | yes |
| cardiac\_index | Cardiac index | CI |  |  | yes |
| mRAP | Mean right atrial pressure | mRAP | mmHg |  | yes |
| WHOFc\_class | WHO Functional Classification | WHO class |  | I/II; III/IV | yes |
| SO2\_RL\_class | O2 saturation | SO2 | % | ≥95; <95 | yes |
| mRASP | mRASP risk score | mRASP |  |  | no |
| Compera | COMPERA score | COMPERA |  |  | no |
| SPAHR | SAPHR score | SPAHR |  |  | no |
| FRENCH3p | FRENCH score, 3 parameters | FPHR 3p |  |  | no |
| FRENCH4p | FRENCH score, 4 parameters | FPHR 4p |  |  | no |
| Reveal\_lite2\_3\_cat | Reveal Lite, Risk Classes | Reveal Lite |  |  | no |
| Reveal2\_risk\_3\_cat | Reveal 2.0 Risk classes | Reveal 2.0 |  |  | no |
| 1Variable name in the R analysis pipeline. | | | | | |
| 2Variable name in the figures and tables. | | | | | |

**Supplementary** **Table S2:** Supplementary characteristic of the study cohorts.

| **Variable** | **IBK** | **LZ/W** | **Significance** | **Effect size** |
| --- | --- | --- | --- | --- |
| N participants | 100 | 83 |  |  |
| mRAP, mmHg | median: 10 [IQR: 6 - 13] range: 2 - 26 | median: 6 [IQR: 3 - 9] range: 0 - 20 | p < 0.0011 | r = 0.431 |
| SO2, % | ≥95: 47% (n = 47) <95: 53% (n = 53) | ≥95: 48% (n = 40) <95: 52% (n = 43) | ns (p = 0.99)3 | V = 0.0124 |
| log NT-pro-BNP, pg/ml | median: 6.6 [IQR: 5.1 - 7.7] range: 3.4 - 11 | median: 6.6 [IQR: 5.3 - 7.4] range: 3.2 - 10 | ns (p = 0.95)1 | r = 0.00791 |
| CI | median: 2.4 [IQR: 1.9 - 2.8] range: 1.6 - 4.2 | median: 2.6 [IQR: 2.2 - 3] range: 1.4 - 3.8 | p = 0.0331 | r = 0.191 |
| RAA, cm2 | median: 22 [IQR: 17 - 24] range: 13 - 34 | median: 18 [IQR: 17 - 23] range: 13 - 30 | p = 0.0121 | r = 0.221 |
| MCV, fl | median: 88 [IQR: 85 - 91] range: 58 - 100 | median: 89 [IQR: 86 - 93] range: 76 - 110 | ns (p = 0.3)1 | r = 0.11 |
| log RDW, % | median: 2.7 [IQR: 2.6 - 2.8] range: 2.5 - 3.1 | median: 2.7 [IQR: 2.7 - 2.8] range: 2.5 - 3.1 | p = 0.0121 | r = 0.221 |
| log FT, ng/ml | median: 4.1 [IQR: 3.5 - 4.7] range: 1.1 - 6.5 | median: 4.4 [IQR: 3.4 - 4.9] range: 1.9 - 7.1 | ns (p = 0.58)1 | r = 0.0561 |
| log TF-Sat, % | median: 3 [IQR: 2.6 - 3.3] range: 0.69 - 4.3 | median: 3 [IQR: 2.5 - 3.4] range: 0.69 - 4.5 | ns (p = 0.5)1 | r = 0.0661 |
| mRASP | median: 1 [IQR: 0 - 1] range: 0 - 2 | median: 1 [IQR: 0 - 1] range: 0 - 2 | ns (p = 0.3)1 | r = 0.11 |
| COMPERA | median: 2 [IQR: 2 - 2] range: 1 - 3 | median: 2 [IQR: 1 - 2] range: 1 - 3 | p = 0.00911 | r = 0.231 |
| SPAHR | median: 2 [IQR: 2 - 2] range: 1 - 3 | median: 2 [IQR: 1 - 2] range: 1 - 3 | ns (p = 0.065)1 | r = 0.171 |
| FPHR 3p | median: 2 [IQR: 1 - 3] range: 0 - 3 | median: 2 [IQR: 1 - 3] range: 0 - 3 | ns (p = 0.3)1 | r = 0.0991 |
| FPHR 4p | median: 3 [IQR: 2 - 4] range: 0 - 4 | median: 2 [IQR: 1 - 3] range: 0 - 4 | p < 0.0011 | r = 0.331 |
| 3-year mortality | 13% (n = 13) | 11% (n = 9) | ns (p = 0.94)3 | V = 0.0334 |
| Overall mortality | 33% (n = 33) | 24% (n = 20) | ns (p = 0.35)3 | V = 0.0984 |
| 1Mann-Whitney U test.1 | | | | |
| 1Wilcoxon r effect size statistic.2 | | | | |
| 3χ² test. | | | | |
| 4Cramer V effect size statistic. | | | | |

**Supplementary** **Table S3:** Results of univariable Cox modeling.

| **Cohort** | **Variable** | **Level** | **Model order** | **HR1** | **Significance** | **C index2** | **R²** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| IBK | Age |  | 1 | 2.7 [1.3 - 5.3] | p = 0.016 | 0.68 [0.59 - 0.78] | 0.33 |
|  | 2 | 0.71 [0.32 - 1.5] | ns (p = 0.47) | 0.68 [0.59 - 0.78] | 0.33 |
| SMWD |  | 1 | 0.41 [0.24 - 0.7] | p = 0.0068 | 0.7 [0.59 - 0.8] | 0.3 |
|  | 2 | 0.82 [0.56 - 1.2] | ns (p = 0.4) | 0.7 [0.59 - 0.8] | 0.3 |
| mPAP |  | 1 | 4.6 [2.2 - 9.4] | p < 0.001 | 0.74 [0.65 - 0.83] | 0.38 |
|  | 2 | 0.63 [0.45 - 0.88] | p = 0.02 | 0.74 [0.65 - 0.83] | 0.38 |
| Sex | male |  | 2 [1 - 4] | ns (p = 0.12) | 0.61 [0.52 - 0.7] | 0.069 |
| PVR |  | 1 | 3.5 [1.6 - 7.8] | p = 0.0095 | 0.68 [0.59 - 0.78] | 0.25 |
|  | 2 | 0.73 [0.52 - 1] | ns (p = 0.15) | 0.68 [0.59 - 0.78] | 0.25 |
| PCWP |  | 1 | 1.3 [0.76 - 2.2] | ns (p = 0.45) | 0.51 [0.41 - 0.61] | 0.017 |
|  | 2 | 0.92 [0.71 - 1.2] | ns (p = 0.59) | 0.51 [0.41 - 0.61] | 0.017 |
| Anemia | yes |  | 1.4 [0.59 - 3.2] | ns (p = 0.55) | 0.54 [0.46 - 0.62] | 0.0093 |
| log RDW |  | 1 | 1.8 [0.96 - 3.3] | ns (p = 0.14) | 0.66 [0.55 - 0.77] | 0.2 |
|  | 2 | 0.99 [0.75 - 1.3] | ns (p = 0.92) | 0.66 [0.55 - 0.77] | 0.2 |
| Renal insufficiency | yes |  | 2.5 [1.3 - 5] | p = 0.021 | 0.6 [0.51 - 0.69] | 0.13 |
| log FT |  | 1 | 1.3 [0.88 - 1.9] | ns (p = 0.3) | 0.6 [0.49 - 0.71] | 0.039 |
|  | 2 | 0.93 [0.72 - 1.2] | ns (p = 0.59) | 0.6 [0.49 - 0.71] | 0.039 |
| log TF-Sat |  | 1 | 1 [0.7 - 1.5] | ns (p = 0.9) | 0.58 [0.48 - 0.68] | 0.056 |
|  | 2 | 1.2 [0.97 - 1.4] | ns (p = 0.16) | 0.58 [0.48 - 0.68] | 0.056 |
| MCV |  | 1 | 1.4 [1 - 2.1] | ns (p = 0.12) | 0.54 [0.41 - 0.66] | 0.13 |
|  | 2 | 1.2 [1.1 - 1.3] | p = 0.0087 | 0.54 [0.41 - 0.66] | 0.13 |
| log NT-pro-BNP |  | 1 | 3.9 [2 - 7.7] | p < 0.001 | 0.76 [0.67 - 0.85] | 0.49 |
|  | 2 | 0.58 [0.4 - 0.84] | p = 0.013 | 0.76 [0.67 - 0.85] | 0.49 |
| Percardial effusion | yes |  | 2.1 [0.91 - 4.9] | ns (p = 0.15) | 0.56 [0.48 - 0.63] | 0.048 |
| RAA |  | 1 | 3.6 [1.8 - 7.2] | p = 0.0034 | 0.73 [0.65 - 0.81] | 0.43 |
|  | 2 | 0.46 [0.29 - 0.74] | p = 0.0068 | 0.73 [0.65 - 0.81] | 0.43 |
| CI |  | 1 | 0.31 [0.19 - 0.52] | p < 0.001 | 0.77 [0.68 - 0.85] | 0.4 |
|  | 2 | 1.3 [0.84 - 2] | ns (p = 0.35) | 0.77 [0.68 - 0.85] | 0.4 |
| mRAP |  | 1 | 1.5 [0.93 - 2.5] | ns (p = 0.16) | 0.56 [0.46 - 0.66] | 0.074 |
|  | 2 | 0.92 [0.72 - 1.2] | ns (p = 0.59) | 0.56 [0.46 - 0.66] | 0.074 |
| WHO class | III/IV |  | 1.9 [0.86 - 4] | ns (p = 0.18) | 0.55 [0.46 - 0.64] | 0.05 |
| SO2 | <95 |  | 1.5 [0.76 - 3.1] | ns (p = 0.34) | 0.58 [0.49 - 0.67] | 0.027 |
| LZ/W | Age |  | 1 | 3.8 [1.2 - 12] | ns (p = 0.076) | 0.69 [0.56 - 0.82] | 0.26 |
|  | 2 | 1.4 [0.78 - 2.5] | ns (p = 0.43) | 0.69 [0.56 - 0.82] | 0.26 |
| SMWD |  | 1 | 0.44 [0.24 - 0.8] | p = 0.048 | 0.68 [0.56 - 0.8] | 0.3 |
|  | 2 | 0.93 [0.62 - 1.4] | ns (p = 0.79) | 0.68 [0.56 - 0.8] | 0.3 |
| mPAP |  | 1 | 1.9 [1.1 - 3.4] | ns (p = 0.086) | 0.63 [0.5 - 0.76] | 0.19 |
|  | 2 | 0.81 [0.53 - 1.2] | ns (p = 0.49) | 0.63 [0.5 - 0.76] | 0.19 |
| Sex | male |  | 5.7 [2.2 - 15] | p = 0.011 | 0.73 [0.64 - 0.83] | 0.33 |
| PVR |  | 1 | 5.8 [2 - 16] | p = 0.011 | 0.74 [0.61 - 0.87] | 0.38 |
|  | 2 | 0.54 [0.33 - 0.87] | ns (p = 0.064) | 0.74 [0.61 - 0.87] | 0.38 |
| PCWP |  | 1 | 0.73 [0.31 - 1.7] | ns (p = 0.55) | 0.61 [0.46 - 0.76] | 0.17 |
|  | 2 | 0.48 [0.2 - 1.1] | ns (p = 0.25) | 0.61 [0.46 - 0.76] | 0.17 |
| Anemia | yes |  | 1.9 [0.69 - 5.4] | ns (p = 0.4) | 0.52 [0.43 - 0.6] | 0.042 |
| log RDW |  | 1 | 1.7 [0.91 - 3] | ns (p = 0.25) | 0.61 [0.47 - 0.74] | 0.1 |
|  | 2 | 0.93 [0.69 - 1.2] | ns (p = 0.68) | 0.61 [0.47 - 0.74] | 0.1 |
| Renal insufficiency | yes |  | 1.5 [0.49 - 4.5] | ns (p = 0.55) | 0.49 [0.41 - 0.57] | 0.014 |
| log FT |  | 1 | 1.4 [0.93 - 2.1] | ns (p = 0.26) | 0.58 [0.43 - 0.72] | 0.075 |
|  | 2 | 1.1 [0.88 - 1.4] | ns (p = 0.54) | 0.58 [0.43 - 0.72] | 0.075 |
| log TF-Sat |  | 1 | 0.61 [0.34 - 1.1] | ns (p = 0.25) | 0.61 [0.47 - 0.75] | 0.097 |
|  | 2 | 0.86 [0.59 - 1.3] | ns (p = 0.55) | 0.61 [0.47 - 0.75] | 0.097 |
| MCV |  | 1 | 1.4 [0.82 - 2.5] | ns (p = 0.4) | 0.62 [0.46 - 0.78] | 0.059 |
|  | 2 | 0.84 [0.54 - 1.3] | ns (p = 0.55) | 0.62 [0.46 - 0.78] | 0.059 |
| log NT-pro-BNP |  | 1 | 2.2 [1.3 - 3.6] | p = 0.02 | 0.67 [0.52 - 0.82] | 0.31 |
|  | 2 | 0.96 [0.69 - 1.3] | ns (p = 0.83) | 0.67 [0.52 - 0.82] | 0.31 |
| Percardial effusion | yes |  | 4 [0.89 - 18] | ns (p = 0.22) | 0.53 [0.47 - 0.59] | 0.071 |
| RAA |  | 1 | 1.5 [0.66 - 3.5] | ns (p = 0.49) | 0.65 [0.53 - 0.78] | 0.12 |
|  | 2 | 1 [0.62 - 1.7] | ns (p = 0.95) | 0.65 [0.53 - 0.78] | 0.12 |
| CI |  | 1 | 0.63 [0.35 - 1.1] | ns (p = 0.27) | 0.68 [0.55 - 0.82] | 0.087 |
|  | 2 | 0.84 [0.56 - 1.3] | ns (p = 0.54) | 0.68 [0.55 - 0.82] | 0.087 |
| mRAP |  | 1 | 2.4 [1.2 - 4.8] | ns (p = 0.072) | 0.72 [0.59 - 0.84] | 0.25 |
|  | 2 | 0.79 [0.52 - 1.2] | ns (p = 0.43) | 0.72 [0.59 - 0.84] | 0.25 |
| WHO class | III/IV |  | 5.5 [2 - 15] | p = 0.011 | 0.69 [0.58 - 0.79] | 0.36 |
| SO2 | <95 |  | 1.7 [0.71 - 4.3] | ns (p = 0.4) | 0.56 [0.43 - 0.68] | 0.045 |
| 1Hazard ratio with 95% confidence interval. | | | | | | | |
| 2Concordance index with 95% confidence interval. | | | | | | | |

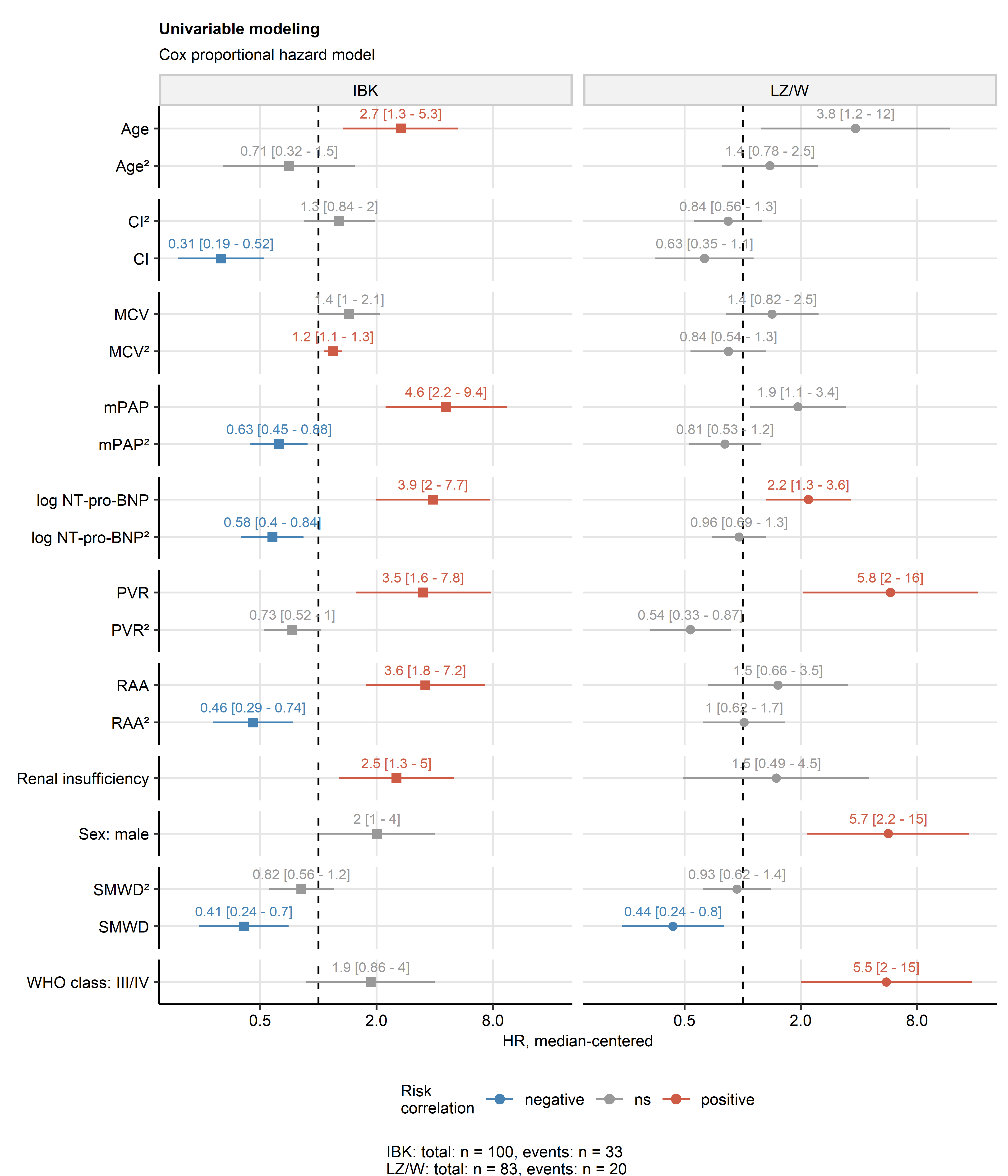
**Supplementary** **Table S4:** Characteristic of the participant clusters in the Innsbruck cohort.

| **Variable** | **Cluster #1** | **Cluster #2** | **Significance** | **Effect size** |
| --- | --- | --- | --- | --- |
| N participants | 46 | 54 |  |  |
| Age, y | median: 58 [IQR: 48 - 68] range: 4 - 80 | median: 69 [IQR: 65 - 74] range: 26 - 84 | p = 0.00151 | r = 0.341 |
| SMWD, m | median: 370 [IQR: 310 - 450] range: 120 - 580 | median: 240 [IQR: 160 - 330] range: 50 - 610 | p < 0.0011 | r = 0.471 |
| mPAP, mmHg | median: 30 [IQR: 27 - 37] range: 26 - 87 | median: 47 [IQR: 40 - 55] range: 26 - 120 | p < 0.0013 | r = 0.564 |
| Sex | female: 78% (n = 36) male: 22% (n = 10) | female: 52% (n = 28) male: 48% (n = 26) | p = 0.0181 | V = 0.271 |
| PVR, Wood | median: 6.8 [IQR: 5.9 - 9.6] range: 3.3 - 38 | median: 14 [IQR: 11 - 22] range: 4.3 - 43 | p < 0.0011 | r = 0.61 |
| Anemia | 15% (n = 7) | 22% (n = 12) | ns (p = 0.53)3 | V = 0.0894 |
| log RDW, % | median: 2.6 [IQR: 2.6 - 2.7] range: 2.5 - 2.9 | median: 2.7 [IQR: 2.6 - 2.8] range: 2.6 - 3.1 | p < 0.0011 | r = 0.381 |
| Renal insufficiency | 17% (n = 8) | 50% (n = 27) | p = 0.00263 | V = 0.344 |
| log FT, ng/ml | median: 4 [IQR: 3.3 - 4.7] range: 1.1 - 6.5 | median: 4.2 [IQR: 3.8 - 4.9] range: 1.8 - 6.5 | ns (p = 0.23)1 | r = 0.121 |
| log TF-Sat, % | median: 3 [IQR: 2.7 - 3.4] range: 1.8 - 4.3 | median: 2.8 [IQR: 2.4 - 3.2] range: 0.69 - 4 | ns (p = 0.16)1 | r = 0.151 |
| MCV, fl | median: 88 [IQR: 85 - 90] range: 76 - 96 | median: 88 [IQR: 86 - 92] range: 58 - 100 | ns (p = 0.33)1 | r = 0.11 |
| log NT-pro-BNP, pg/ml | median: 5.1 [IQR: 4.4 - 5.8] range: 3.4 - 7.5 | median: 7.6 [IQR: 6.9 - 8.1] range: 4.9 - 11 | p < 0.0011 | r = 0.781 |
| Percardial effusion | 6.5% (n = 3) | 24% (n = 13) | p = 0.0473 | V = 0.244 |
| RAA, cm² | median: 17 [IQR: 16 - 21] range: 13 - 27 | median: 24 [IQR: 23 - 27] range: 15 - 34 | p < 0.0011 | r = 0.671 |
| CI | median: 2.7 [IQR: 2.4 - 3] range: 1.8 - 4.2 | median: 2 [IQR: 1.9 - 2.3] range: 1.6 - 3.5 | p < 0.0011 | r = 0.61 |
| mRAP, mmHg | median: 8 [IQR: 6 - 12] range: 2 - 18 | median: 11 [IQR: 8 - 14] range: 2 - 26 | p = 0.0241 | r = 0.241 |
| WHO class | I/II: 48% (n = 22) III/IV: 52% (n = 24) | I/II: 31% (n = 17) III/IV: 69% (n = 37) | ns (p = 0.16)3 | V = 0.174 |
| SO2, % | ≥95: 57% (n = 26) <95: 43% (n = 20) | ≥95: 39% (n = 21) <95: 61% (n = 33) | ns (p = 0.15)1 | V = 0.181 |
| 3-year mortality | 6.5% (n = 3) | 19% (n = 10) | ns (p = 0.16)3 | V = 0.184 |
| 5-year mortality | 6.5% (n = 3) | 33% (n = 18) | p = 0.00423 | V = 0.334 |
| mRASP | median: 0 [IQR: 0 - 1] range: 0 - 1 | median: 1 [IQR: 1 - 2] range: 0 - 2 | p < 0.0011 | r = 0.681 |
| COMPERA | median: 2 [IQR: 1 - 2] range: 1 - 2 | median: 2 [IQR: 2 - 2.8] range: 1 - 3 | p < 0.0011 | r = 0.581 |
| SPAHR | median: 1.5 [IQR: 1 - 2] range: 1 - 2 | median: 2 [IQR: 2 - 2] range: 1 - 3 | p < 0.0011 | r = 0.591 |
| FPHR 3p | median: 2 [IQR: 1 - 2] range: 0 - 3 | median: 3 [IQR: 2 - 3] range: 1 - 3 | p < 0.0011 | r = 0.551 |
| FPHR 4p | median: 2 [IQR: 2 - 3] range: 0 - 4 | median: 3 [IQR: 3 - 4] range: 1 - 4 | p < 0.0011 | r = 0.491 |
| 1Mann-Whitney U test.1 | | | | |
| 1Wilcoxon r effect size statistic.2 | | | | |
| 3χ² test. | | | | |
| 4Cramer V effect size statistic. | | | | |

**Supplementary** **Table S5:** Characteristic of the participant clusters in the Linz/Vienna cohort.

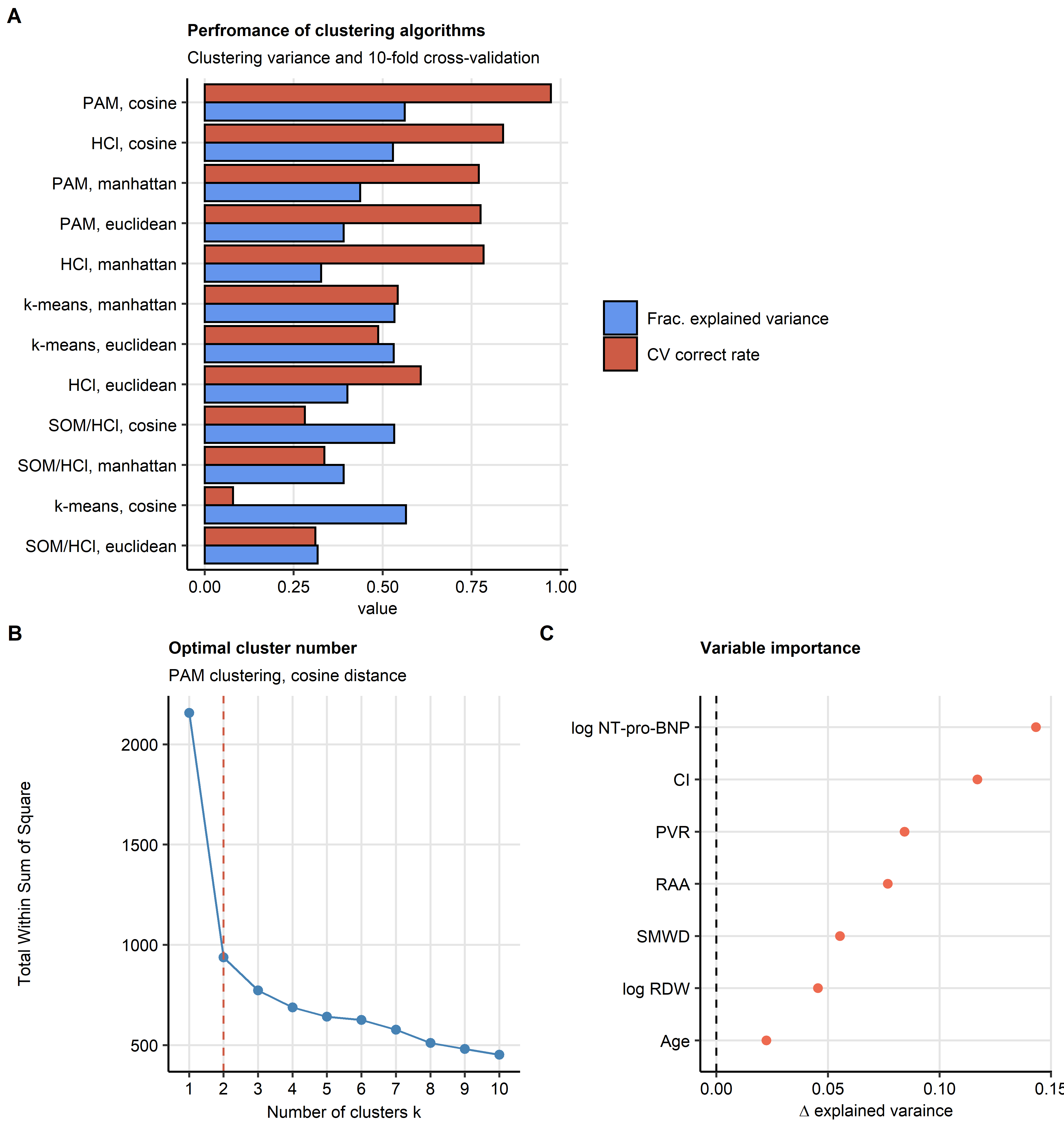
| **Variable** | **Cluster #1** | **Cluster #2** | **Significance** | **Effect size** |
| --- | --- | --- | --- | --- |
| N participants | 35 | 48 |  |  |
| Age, y | median: 63 [IQR: 46 - 71] range: 23 - 81 | median: 71 [IQR: 64 - 74] range: 26 - 82 | p = 0.00971 | r = 0.311 |
| SMWD, m | median: 430 [IQR: 350 - 510] range: 190 - 620 | median: 320 [IQR: 180 - 380] range: 50 - 580 | p < 0.0011 | r = 0.461 |
| mPAP, mmHg | median: 34 [IQR: 28 - 39] range: 18 - 57 | median: 44 [IQR: 35 - 50] range: 28 - 67 | p < 0.0013 | r = 0.474 |
| Sex | female: 77% (n = 27) male: 23% (n = 8) | female: 58% (n = 28) male: 42% (n = 20) | ns (p = 0.16)1 | V = 0.21 |
| PVR, Wood | median: 3.7 [IQR: 3 - 5] range: 1.4 - 10 | median: 6.4 [IQR: 4.9 - 9] range: 2.3 - 20 | p < 0.0011 | r = 0.461 |
| Anemia | 5.7% (n = 2) | 25% (n = 12) | ns (p = 0.07)3 | V = 0.254 |
| log RDW, % | median: 2.7 [IQR: 2.6 - 2.8] range: 2.5 - 2.8 | median: 2.8 [IQR: 2.7 - 2.8] range: 2.5 - 3.1 | p = 0.0111 | r = 0.31 |
| Renal insufficiency | 11% (n = 4) | 23% (n = 11) | ns (p = 0.33)3 | V = 0.154 |
| log FT, ng/ml | median: 4.2 [IQR: 3 - 4.8] range: 1.9 - 5.7 | median: 4.4 [IQR: 3.6 - 5] range: 2.5 - 7.1 | ns (p = 0.12)1 | r = 0.191 |
| log TF-Sat, % | median: 3.1 [IQR: 2.8 - 3.4] range: 1.6 - 4.5 | median: 2.8 [IQR: 2.4 - 3.4] range: 0.69 - 4.1 | ns (p = 0.16)1 | r = 0.171 |
| MCV, fl | median: 89 [IQR: 84 - 94] range: 78 - 110 | median: 90 [IQR: 87 - 93] range: 76 - 100 | ns (p = 0.68)1 | r = 0.0451 |
| log NT-pro-BNP, pg/ml | median: 5.3 [IQR: 4.9 - 6] range: 3.2 - 7 | median: 7.3 [IQR: 6.7 - 7.9] range: 4.8 - 10 | p < 0.0011 | r = 0.681 |
| Percardial effusion | 0% (n = 0) | 6.2% (n = 3) | ns (p = 0.38)3 | V = 0.174 |
| RAA, cm² | median: 17 [IQR: 15 - 17] range: 13 - 20 | median: 22 [IQR: 19 - 25] range: 15 - 30 | p < 0.0011 | r = 0.691 |
| CI | median: 2.8 [IQR: 2.5 - 3.1] range: 1.8 - 3.8 | median: 2.4 [IQR: 2.1 - 2.8] range: 1.4 - 3.6 | p = 0.00971 | r = 0.311 |
| mRAP, mmHg | median: 3 [IQR: 1 - 5.5] range: 0 - 16 | median: 8 [IQR: 5 - 9.2] range: 1 - 20 | p < 0.0011 | r = 0.551 |
| WHO class | I/II: 80% (n = 28) III/IV: 20% (n = 7) | I/II: 33% (n = 16) III/IV: 67% (n = 32) | p < 0.0013 | V = 0.464 |
| SO2, % | ≥95: 60% (n = 21) <95: 40% (n = 14) | ≥95: 40% (n = 19) <95: 60% (n = 29) | ns (p = 0.15)1 | V = 0.21 |
| 3-year mortality | 5.7% (n = 2) | 15% (n = 7) | ns (p = 0.38)3 | V = 0.144 |
| 5-year mortality | 5.7% (n = 2) | 19% (n = 9) | ns (p = 0.19)3 | V = 0.194 |
| mRASP | median: 0 [IQR: 0 - 0] range: 0 - 1 | median: 1 [IQR: 1 - 1] range: 0 - 2 | p < 0.0011 | r = 0.81 |
| COMPERA | median: 1 [IQR: 1 - 1] range: 1 - 2 | median: 2 [IQR: 2 - 2] range: 1 - 3 | p < 0.0011 | r = 0.721 |
| SPAHR | median: 1 [IQR: 1 - 1] range: 1 - 2 | median: 2 [IQR: 2 - 2] range: 1 - 3 | p < 0.0011 | r = 0.781 |
| FPHR 3p | median: 1 [IQR: 0 - 2] range: 0 - 3 | median: 3 [IQR: 2 - 3] range: 1 - 3 | p < 0.0011 | r = 0.621 |
| FPHR 4p | median: 1 [IQR: 0 - 2] range: 0 - 3 | median: 3 [IQR: 2 - 3] range: 1 - 4 | p < 0.0011 | r = 0.61 |
| 1Mann-Whitney U test.1 | | | | |
| 1Wilcoxon r effect size statistic.2 | | | | |
| 3χ² test. | | | | |
| 4Cramer V effect size statistic. | | | | |

# Supplementary Figures



**Supplementary Figure S1. Univariable Cox proportional hazard modeling.**

Association of candidate risk factors (**Supplementary Table S1**) with overall survival was investigated with a series of univariable Cox proportional hazard models. Numeric independent variables were median-centered and their first and second order terms included in the models. Hazard ratio (HR) estimate significance was determined by Wald Z test and adjusted for multiple testing with Benjamini-Hochberg method. HR values with 95 confidence intervals for variables significantly associated with the survival in at least one Innsbruck (IBK) or Linz/Vienna cohort (LZ/W) were presented in a Forest plot. Numbers of complete observations and mortality are indicated under the plot. CI: cardiac index; MCV: mean corpuscular volume; mPAP: mean pulmonary artherial pressure; NT-pro-BNP: N terminal pro brain natriuretic peptide; PVR: pulmonary vascular resistance; RAA: right atrial area; SMWD: six minute walking distance.



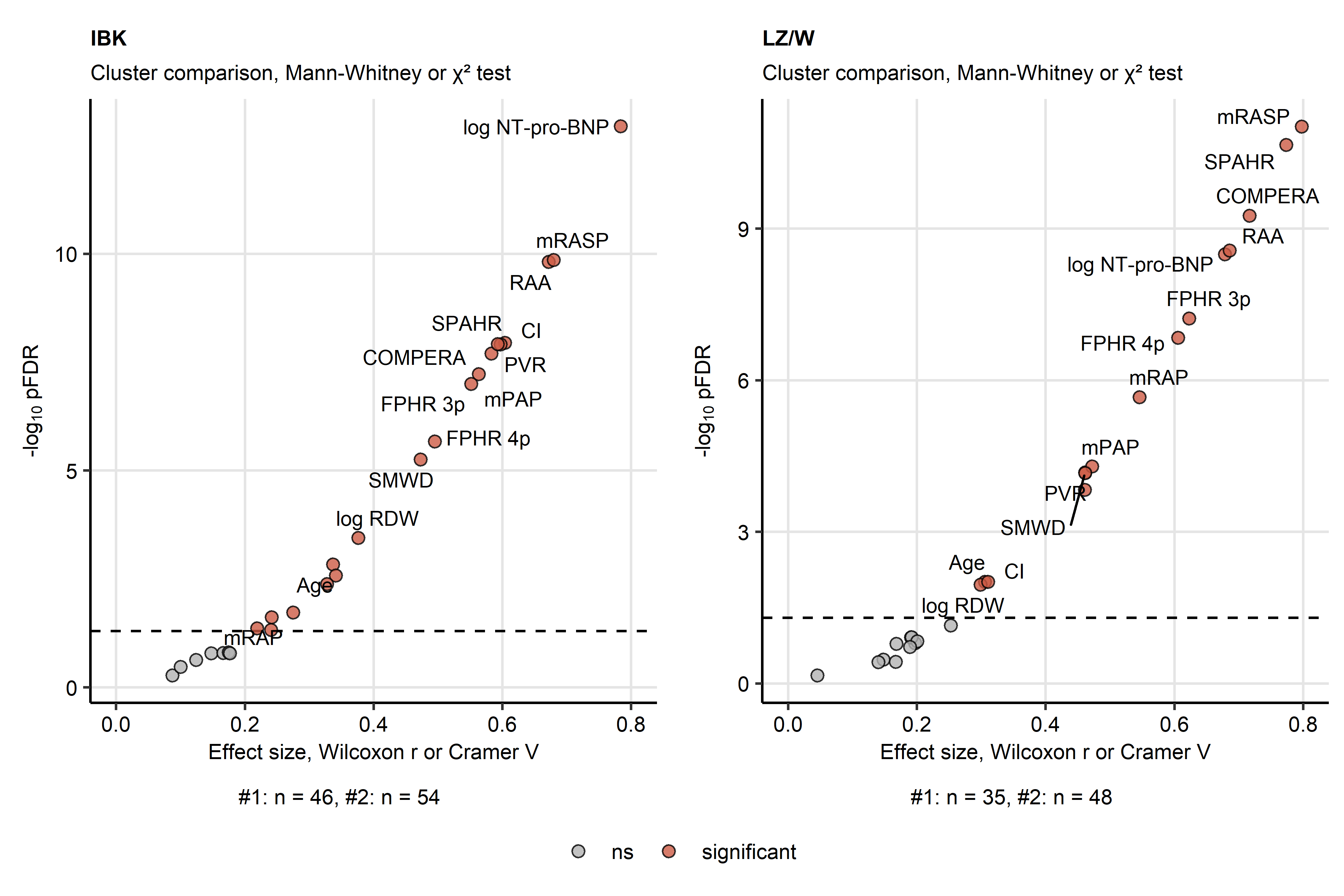
**Supplementary Figure S2. Development of participant clusters.**

Clustering of the training Innsbruck (IBK) cohort participants in respect to the survival-associated factors identified by elastic-net modeling (**Figure 2A**) was investigated by PAM (partitioning around medoids) algorithm and cosine distance measure.

**(A)** Comparison of the ‘explained’ clustering variance (between-cluster to total sum-of-squares) and 10-fold cross-validation (CV) correct prediction rate for clustering of the IBK cohort with various algorithms (PAM, HCl: hierarchical clustering, k-means and SOM/HCl: combined self-organizing map/hierarchical clustering) and distance statistics (Euclidean, Manhattan and cosine distance). Note the superior ‘explained’ variance fraction and CV performance of the PAM algorithm/cosine distance procedure.

**(B)** Determination of the optimal cluster number by the bend of the total within-cluster sum-of-squares curve. The dashed vertical line indicates the chosen number of PAM clusters.

**(C)** Importance of particular clustering features was determined by comparing the ‘explained’ clustering variances of the original clustering structure with the clustering objects with randomly re-shuffled clustering features.



**Supplementary Figure S3. Differences in study variable between the participant clusters.**

Training Innsbruck (IBK) cohort participants were clustered as presented in **Figure 4** and **Supplementary Figure S2**. Cluster assignment of the test Linz/Vienna (LZ/W) cohort participants was accomplished by k-nearest neighbor label propagation procedure. Differences in the study variables (**Supplementary Table S1**) between the clusters were determined by Mann-Whitney test with r effect size statistic or by test with Cramer V effect size statistic for numeric and categorical features, respectively. P values were adjusted for multiple testing with Benjamini-Hochberg method (pFDR). Significance (pFDR) and effect size are presented in scatter plots. Each point represents one study parameter, parameters significantly different between the clusters are highlighted in red. Parameters found significant in both cohorts are labeled with their names. The significance cutoff is depicted as a dashed line. Numbers of participants assigned to the clusters are presented under the plots. CI: cardiac index; mPAP: mean pulmonary arterial pressure; NT-pro-BNP: N terminal pro brain natriuretic peptide; PVR: pulmonary vascular resistance; RAA: right atrial area; SMWD: six minute walking distance; mPAP: mean pulmonary arterial pressure; RDW: red blood cell distribution width; mRAP: mean right atrial pressure; FPHR: French pulmonary hypertension register; SPAHR: Swedish pulmonary arterial hypertension register; COMPERA: comparative, prospective registry of newly initiated therapies for pulmonary hypertension; mRASP: modified risk assessment score of PAH.

# References

1. Wickham, H. *et al.* Welcome to the Tidyverse. *Journal of Open Source Software* **4**, 1686 (2019).

2. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer-Verlag, 2016).

3. Wilke, C. O. *Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures*. 389 (O’Reilly Media, 2019).

4. Xie, Y. *Bookdown : authoring books and technical documents with R Markdown*. 113 (2016).

5. Harrington, D. P. & Fleming, T. R. A Class of Rank Test Procedures for Censored Survival Data. *Biometrika* **69**, 553 (1982).

6. Therneau, T. M. & Grambsch, P. M. *Modeling Survival Data: Extending the Cox Model*. (Springer Verlag, 2000).

7. Kassambara, A., Kosinski, M. & Biecek, P. survminer: Drawing Survival Curves using ’ggplot2’. (2016).

8. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 289–300 (1995).

9. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**, 1–22 (2010).

10. Schubert, E. & Rousseeuw, P. J. 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)* vol. 11807 LNCS 171–187 (Springer, 2019).

11. Drost, H.-G. Philentropy: Information Theory and Distance Quantification with R. *Journal of Open Source Software* **3**, 765 (2018).

12. Lange, T., Roth, V., Braun, M. L. & Buhmann, J. M. Stability-Based Validation of Clustering Solutions. *Neural Computation* **16**, 1299–1323 (2004).

13. Sonnweber, T. *et al.* Investigating phenotypes of pulmonary COVID-19 recovery - a longitudinal observational prospective multicenter trial. *eLife* **11**, (2022).

14. Leng, M., Wang, J., Cheng, J., Zhou, H. & Chen, X. Adaptive semi-supervised clustering algorithm with label propagation. *Journal of Software Engineering* **8**, 14–22 (2014).

15. Sahanic, S. *et al.* Phenotyping of acute and persistent COVID-19 features in the outpatient setting: exploratory analysis of an international cross-sectional online survey. *Clinical Infectious Diseases* (2021) doi:[10.1093/CID/CIAB978](https://doi.org/10.1093/CID/CIAB978).