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

Supplementary Material

Innsbruck PAH registry

2023-03-08

Supplementary Methods

Data transformation, visualization, descriptive statistic

Data transformation, analysis and result visualization was accomplished by R version 4.0.5 with *tidyverse* environment[1,2]. 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 χ^2 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 (λ = 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 R^2 (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 assignments[12] 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 χ^2 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.

Variable[a]	Description	Label[b]	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

Variable[a]	Description	Label[b]	Unit	Stratification	Used in risk modeling
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

[a]Variable name in the R analysis pipeline.

[b]Variable 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	$p < 0.001$[a]	$r = 0.43$[b]
SO2, %	≥95: 47% (n = 47) <95: 53% (n = 53)	≥95: 48% (n = 40) <95: 52% (n = 43)	ns ($p = 0.99$)[c]	$V = 0.012$[d]
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 ($p = 0.95$)[a]	$r = 0.0079$[b]
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	$p = 0.033$[a]	$r = 0.19$[b]
RAA, cm2	median: 22 [IQR: 17 - 24] range: 13 - 34	median: 18 [IQR: 17 - 23] range: 13 - 30	$p = 0.012$[a]	$r = 0.22$[b]
MCV, fl	median: 88 [IQR: 85 - 91] range: 58 - 100	median: 89 [IQR: 86 - 93] range: 76 - 110	ns ($p = 0.3$)[a]	$r = 0.1$[b]
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	$p = 0.012$[a]	$r = 0.22$[b]
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 ($p = 0.58$)[a]	$r = 0.056$[b]
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 ($p = 0.5$)[a]	$r = 0.066$[b]
mRASP	median: 1 [IQR: 0 - 1] range: 0 - 2	median: 1 [IQR: 0 - 1] range: 0 - 2	ns ($p = 0.3$)[a]	$r = 0.1$[b]
COMPERA	median: 2 [IQR: 2 - 2] range: 1 - 3	median: 2 [IQR: 1 - 2] range: 1 - 3	$p = 0.0091$[a]	$r = 0.23$[b]
SPAHR	median: 2 [IQR: 2 - 2] range: 1 - 3	median: 2 [IQR: 1 - 2] range: 1 - 3	ns ($p = 0.065$)[a]	$r = 0.17$[b]
FPHR 3p	median: 2 [IQR: 1 - 3] range: 0 - 3	median: 2 [IQR: 1 - 3] range: 0 - 3	ns ($p = 0.3$)[a]	$r = 0.099$[b]
FPHR 4p	median: 3 [IQR: 2 - 4] range: 0 - 4	median: 2 [IQR: 1 - 3] range: 0 - 4	$p < 0.001$[a]	$r = 0.33$[b]
3-year mortality	13% (n = 13)	11% (n = 9)	ns ($p = 0.94$)[c]	$V = 0.033$[d]
Overall mortality	33% (n = 33)	24% (n = 20)	ns ($p = 0.35$)[c]	$V = 0.098$[d]

Variable	IBK	LZ/W	Significance	Effect size
[a]Mann-Whitney U test.				
[b]Wilcoxon r effect size statistic.				
[c]χ^2 test.				
[d]Cramer V effect size statistic.				

Supplementary Table S3: Results of univariable Cox modeling.

Cohort	Variable	Level	Model order	HR[a]	Significance	C index[b]	R^2
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

Cohort	Variable	Level	Model order	HR[a]	Significance	C index[b]	R^2
	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

Cohort	Variable	Level	Model order	HR[a]	Significance	C index[b]	R^2
	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

[a]Hazard ratio with 95% confidence interval.

[b]Concordance 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 = 0.0015[a]	r = 0.34[b]
SMWD, m	median: 370 [IQR: 310 - 450] range: 120 - 580	median: 240 [IQR: 160 - 330] range: 50 - 610	p < 0.001[a]	r = 0.47[b]
mPAP, mmHg	median: 30 [IQR: 27 - 37] range: 26 - 87	median: 47 [IQR: 40 - 55] range: 26 - 120	p < 0.001[c]	r = 0.56[d]
Sex	female: 78% (n = 36) male: 22% (n = 10)	female: 52% (n = 28) male: 48% (n = 26)	p = 0.018[a]	V = 0.27[b]
PVR, Wood	median: 6.8 [IQR: 5.9 - 9.6] range: 3.3 - 38	median: 14 [IQR: 11 - 22] range: 4.3 - 43	p < 0.001[a]	r = 0.6[b]
Anemia	15% (n = 7)	22% (n = 12)	ns (p = 0.53)[c]	V = 0.089[d]
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 < 0.001[a]	r = 0.38[b]
Renal insufficiency	17% (n = 8)	50% (n = 27)	p = 0.0026[c]	V = 0.34[d]
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)[a]	r = 0.12[b]
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)[a]	r = 0.15[b]
MCV, fl	median: 88 [IQR: 85 - 90] range: 76 - 96	median: 88 [IQR: 86 - 92] range: 58 - 100	ns (p = 0.33)[a]	r = 0.1[b]
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 < 0.001[a]	r = 0.78[b]
Percardial effusion	6.5% (n = 3)	24% (n = 13)	p = 0.047[c]	V = 0.24[d]
RAA, cm^2	median: 17 [IQR: 16 - 21] range: 13 - 27	median: 24 [IQR: 23 - 27] range: 15 - 34	p < 0.001[a]	r = 0.67[b]
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 < 0.001[a]	r = 0.6[b]
mRAP, mmHg	median: 8 [IQR: 6 - 12] range: 2 - 18	median: 11 [IQR: 8 - 14] range: 2 - 26	p = 0.024[a]	r = 0.24[b]

Variable	Cluster #1	Cluster #2	Significance	Effect size
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)[c]	V = 0.17[d]
SO2, %	≥95: 57% (n = 26) <95: 43% (n = 20)	≥95: 39% (n = 21) <95: 61% (n = 33)	ns (p = 0.15)[a]	V = 0.18[b]
3-year mortality	6.5% (n = 3)	19% (n = 10)	ns (p = 0.16)[c]	V = 0.18[d]
5-year mortality	6.5% (n = 3)	33% (n = 18)	p = 0.0042[c]	V = 0.33[d]
mRASP	median: 0 [IQR: 0 - 1] range: 0 - 1	median: 1 [IQR: 1 - 2] range: 0 - 2	p < 0.001[a]	r = 0.68[b]
COMPERA	median: 2 [IQR: 1 - 2] range: 1 - 2	median: 2 [IQR: 2 - 2.8] range: 1 - 3	p < 0.001[a]	r = 0.58[b]
SPAHR	median: 1.5 [IQR: 1 - 2] range: 1 - 2	median: 2 [IQR: 2 - 2] range: 1 - 3	p < 0.001[a]	r = 0.59[b]
FPHR 3p	median: 2 [IQR: 1 - 2] range: 0 - 3	median: 3 [IQR: 2 - 3] range: 1 - 3	p < 0.001[a]	r = 0.55[b]
FPHR 4p	median: 2 [IQR: 2 - 3] range: 0 - 4	median: 3 [IQR: 3 - 4] range: 1 - 4	p < 0.001[a]	r = 0.49[b]

[a]Mann-Whitney U test.

[b]r effect size statistic.

[c]χ^2 test.

[d]Cramer 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	$p = 0.0097$[a]	$r = 0.31$[b]
SMWD, m	median: 430 [IQR: 350 - 510] range: 190 - 620	median: 320 [IQR: 180 - 380] range: 50 - 580	$p < 0.001$[a]	$r = 0.46$[b]
mPAP, mmHg	median: 34 [IQR: 28 - 39] range: 18 - 57	median: 44 [IQR: 35 - 50] range: 28 - 67	$p < 0.001$[c]	$r = 0.47$[d]
Sex	female: 77% (n = 27) male: 23% (n = 8)	female: 58% (n = 28) male: 42% (n = 20)	ns ($p = 0.16$)[a]	$V = 0.2$[b]
PVR, Wood	median: 3.7 [IQR: 3 - 5] range: 1.4 - 10	median: 6.4 [IQR: 4.9 - 9] range: 2.3 - 20	$p < 0.001$[a]	$r = 0.46$[b]
Anemia	5.7% (n = 2)	25% (n = 12)	ns ($p = 0.07$)[c]	$V = 0.25$[d]
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	$p = 0.011$[a]	$r = 0.3$[b]
Renal insufficiency	11% (n = 4)	23% (n = 11)	ns ($p = 0.33$)[c]	$V = 0.15$[d]
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 ($p = 0.12$)[a]	$r = 0.19$[b]
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 ($p = 0.16$)[a]	$r = 0.17$[b]
MCV, fl	median: 89 [IQR: 84 - 94] range: 78 - 110	median: 90 [IQR: 87 - 93] range: 76 - 100	ns ($p = 0.68$)[a]	$r = 0.045$[b]
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	$p < 0.001$[a]	$r = 0.68$[b]
Percardial effusion	0% (n = 0)	6.2% (n = 3)	ns ($p = 0.38$)[c]	$V = 0.17$[d]
RAA, cm^2	median: 17 [IQR: 15 - 17] range: 13 - 20	median: 22 [IQR: 19 - 25] range: 15 - 30	$p < 0.001$[a]	$r = 0.69$[b]
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	$p = 0.0097$[a]	$r = 0.31$[b]
mRAP, mmHg	median: 3 [IQR: 1 - 5.5] range: 0 - 16	median: 8 [IQR: 5 - 9.2] range: 1 - 20	$p < 0.001$[a]	$r = 0.55$[b]

Variable	Cluster #1	Cluster #2	Significance	Effect size
WHO class	I/II: 80% (n = 28) III/IV: 20% (n = 7)	I/II: 33% (n = 16) III/IV: 67% (n = 32)	$p < 0.001$[c]	V = 0.46[d]
SO2, %	≥95: 60% (n = 21) <95: 40% (n = 14)	≥95: 40% (n = 19) <95: 60% (n = 29)	ns (p = 0.15)[a]	V = 0.2[b]
3-year mortality	5.7% (n = 2)	15% (n = 7)	ns (p = 0.38)[c]	V = 0.14[d]
5-year mortality	5.7% (n = 2)	19% (n = 9)	ns (p = 0.19)[c]	V = 0.19[d]
mRASP	median: 0 [IQR: 0 - 0] range: 0 - 1	median: 1 [IQR: 1 - 1] range: 0 - 2	$p < 0.001$[a]	r = 0.8[b]
COMPERA	median: 1 [IQR: 1 - 1] range: 1 - 2	median: 2 [IQR: 2 - 2] range: 1 - 3	$p < 0.001$[a]	r = 0.72[b]
SPAHR	median: 1 [IQR: 1 - 1] range: 1 - 2	median: 2 [IQR: 2 - 2] range: 1 - 3	$p < 0.001$[a]	r = 0.78[b]
FPHR 3p	median: 1 [IQR: 0 - 2] range: 0 - 3	median: 3 [IQR: 2 - 3] range: 1 - 3	$p < 0.001$[a]	r = 0.62[b]
FPHR 4p	median: 1 [IQR: 0 - 2] range: 0 - 3	median: 3 [IQR: 2 - 3] range: 1 - 4	$p < 0.001$[a]	r = 0.6[b]

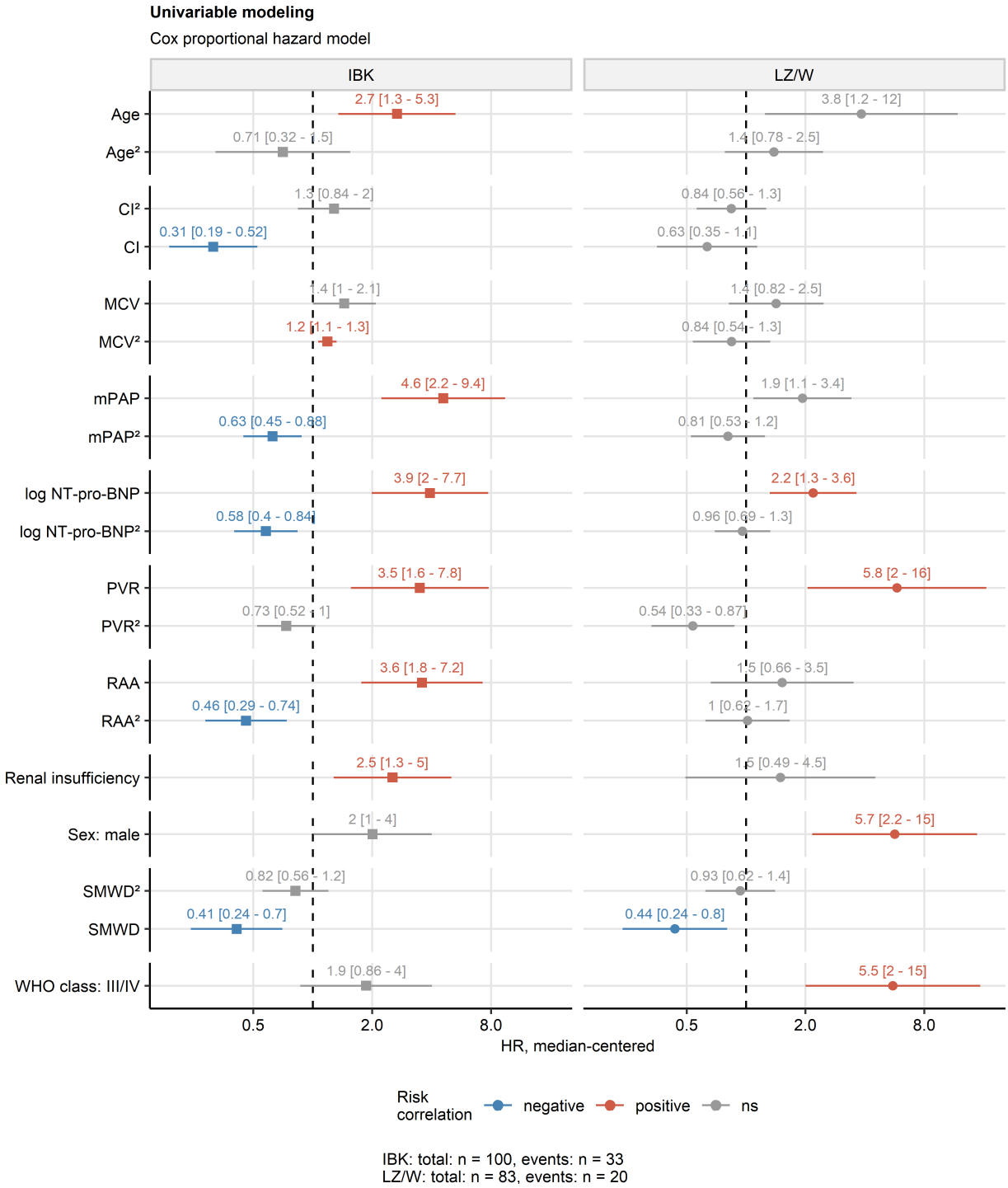
[a]Mann-Whitney U test.

[b]r effect size statistic.

[c]χ^2 test.

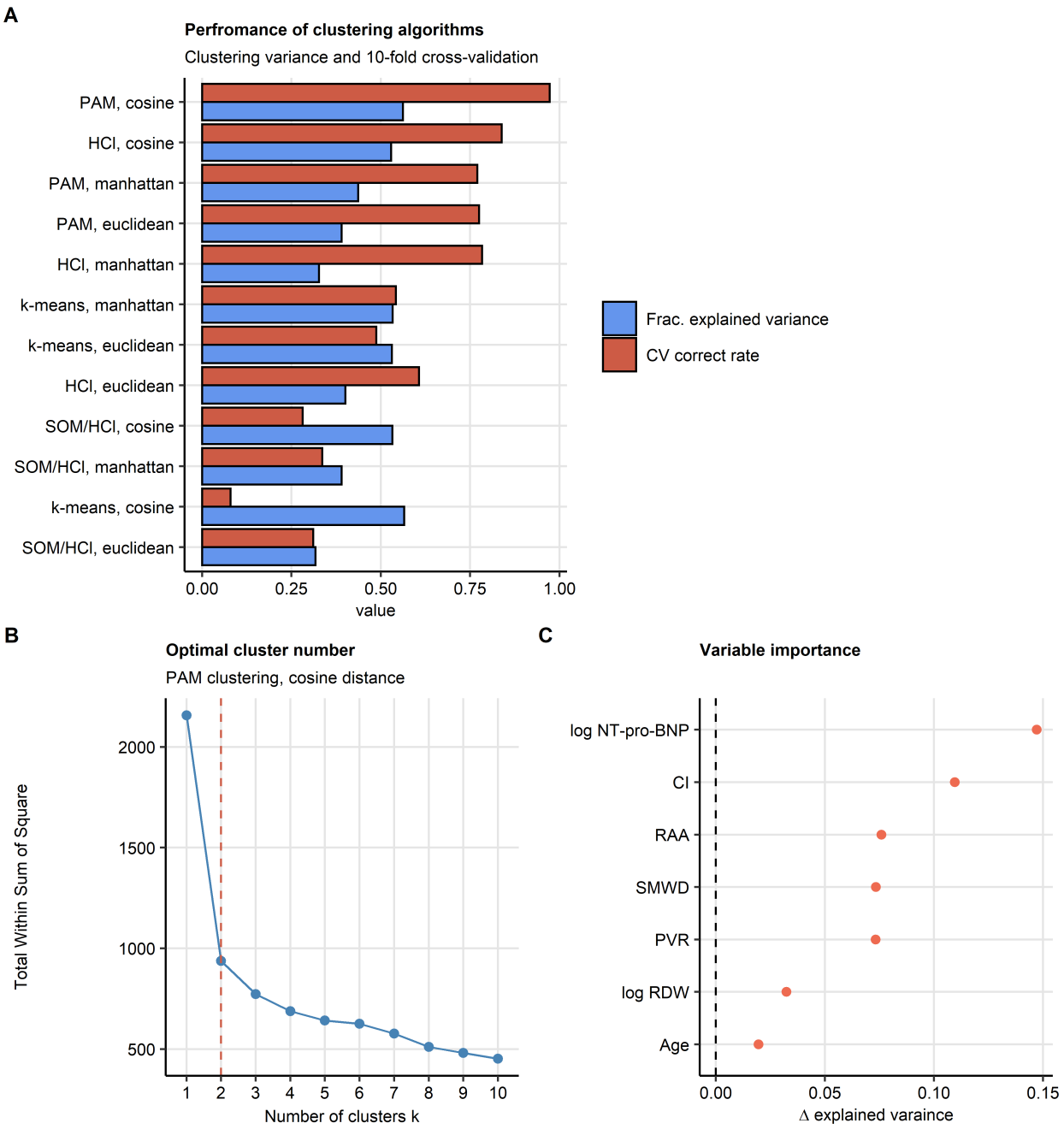
[d]Cramer V effect size statistic.

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.

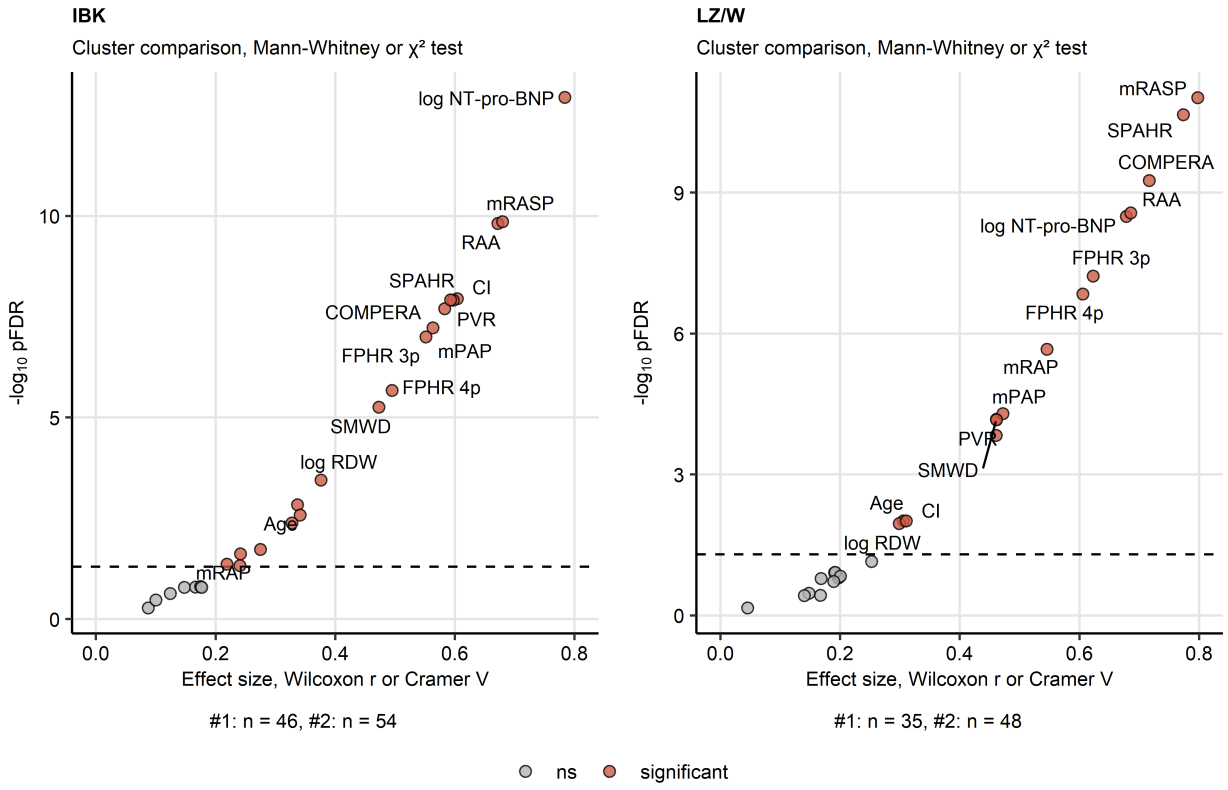
Clustering of the training Innsbruck (IBK) cohort participants in respect to the survival-associated factors identified by elastic-net modeling (**Figure 2A**) was investigated by PAM (partitioning around medoids) algorithm and cosine distance measure.

(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 variables 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 χ^2 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, Averick M, Bryan J, et al. Welcome to the Tidyverse. *Journal of Open Source Software*. 2019;4(43):1686. doi:10.21105/joss.01686

2. Wickham Hadley. *ggplot2: Elegant Graphics for Data Analysis*. 1st ed. Springer-Verlag; 2016. https://ggplot2.tidyverse.org

3. Wilke CO. *Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures*. 1st ed. O'Reilly Media; 2019:389.

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

5. Harrington DP, Fleming TR. A Class of Rank Test Procedures for Censored Survival Data. *Biometrika*. 1982;69(3):553. doi:10.2307/2335991

6. Therneau TM, Grambsch PM. *Modeling Survival Data: Extending the Cox Model*. 1st ed. Springer Verlag; 2000.

7. Kassambara A, Kosinski M, Biecek P. survminer: Drawing Survival Curves using 'ggplot2'. Published online 2016. https://cran.r-project.org/package=survminer

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)*. 1995;57(1):289-300. doi:10.1111/j.2517-6161.1995.tb02031.x

9. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*. 2010;33(1):1-22. doi:10.18637/jss.v033.i01

10. 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)*. Vol 11807 LNCS. Springer; 2019:171-187. doi:10.1007/978-3-030-32047-8_16

11. Drost HG. Philentropy: Information Theory and Distance Quantification with R. *Journal of Open Source Software*. 2018;3(26):765. doi:10.21105/joss.00765

12. Lange T, Roth V, Braun ML, Buhmann JM. Stability-Based Validation of Clustering Solutions. *Neural Computation*. 2004;16(6):1299-1323. doi:10.1162/089976604773717621

13. Sonnweber T, Tymoszuk P, Sahanic S, et al. Investigating phenotypes of pulmonary COVID-19 recovery - a longitudinal observational prospective multicenter trial. *eLife*. 2022;11. doi:10.7554/ELIFE.72500

14. Leng M, Wang J, Cheng J, Zhou H, Chen X. Adaptive semi-supervised clustering algorithm with label propagation. *Journal of Software Engineering*. 2014;8(1):14-22. doi:10.3923/JSE.2014.14.22

15. Sahanic S, Tymoszuk P, Ausserhofer D, 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*. Published online November 2021. doi:10.1093/CID/CIAB978