Online supplement

Identification and validation of distinct biological phenotypes in patients with acute respiratory distress syndrome by cluster analysis.

L.D. Bos, L.R. Schouten, L.A. van Vught, M.A. Wiewel, D. Ong, O. Cremer, A. Artigas, I. Martin-Loeches, A.J. Hoogendijk, T. van der Poll, J. Horn, N. Juffermans, C.S. Calfee, M.J. Schultz
On behalf of the MARS consortium*.*

Members of the MARS consortium

Jos F. Frencken, (Department of Intensive Care Medicine and Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, the Netherlands); Marc Bonten, Peter M. C. Klein Klouwenberg, David Ong (Department of Medical Microbiology, University Medical Center Utrecht, Utrecht, the Netherlands) and Roosmarijn T. M. van Hooijdonk, Mischa A. Huson, Laura R. A. Schouten, Marleen Straat, Lonneke A. van Vught, Maryse A. Wiewel, Esther Witteveen, Gerie J. Glas, and Luuk Wieske, (Department of Intensive Care Medicine, Academic Medical Center, University of Amsterdam); Brendon P Scicluna, Arjan J Hoogendijk, H Belkasim-Bohoudi, Tom van der Poll (Center of Experimental Molecular Medicine; CEMM, Academic Medical Center, University of Amsterdam).

Supplemental methods

Data analysis

Cluster method

Biomarker data was $\log^{10}$ transformed to obtain normally distributed variables, which was checked by Q-Q plots. Hierarchical cluster analysis consists of several steps: dissimilarity matrix, clustering algorithm and estimation of the number of clusters.

The dissimilarity matrix reduces the data from a sample x biomarker matrix to a matrix that describes the "distances" between samples. Several indices are available for these distances. One of the widely used matrices is the Eucledian distance, which is the usual square distance between two vectors. In this case, a vector is the conglomerate of the values for all biomarkers in a given patient. So the Eucledian distance is the square of the distance between the 20-dimensional vector for biomarker concentrations in one patient and a similar vector in another patient. In other words, the Eucledian distance gives a summary value for the (dis)similarity of the biomarker profile between patients.

Ward clustering was performed with this dissimilarity matrix as input. Ward's method minimizes the total within-cluster variance, so it groups the samples with the lowest dissimilarity value (or the highest similarity value). At each step, the pair of clusters with minimum cluster distance is merged. This pair of clusters leads to minimum increase in total within-cluster variance after merging.

As described by Charrad *et al.* [30]: "Most of the clustering algorithms depend on some assumptions in order to define the subgroups present in a data set. As a consequence, the resulting clustering scheme requires some evaluation of its validity. The evaluation procedure has to tackle difficult problems such as the quality of clusters, the degree with which a clustering scheme fits a specific data set and the optimal number of clusters in a partitioning. three approaches to investigate cluster validity are described. The first is based on external criteria, which consist in comparing the results of cluster analysis to externally known results, such as externally provided class labels. The second approach is based on internal criteria, which use the information obtained from within the clustering process to evaluate how well the results of cluster analysis fit the data without reference to external information. The third approach of clustering validity is based on relative criteria, which consists in the evaluation of a clustering structure by comparing it with other clustering schemes, resulting by the same algorithm but with different parameter values, *e.g.*, the number of clusters."

In patients with ARDS, we do not have external criteria to validate the results. So the first method is disqualified. From the two other methods, we choose to evaluate the clustering validity based on relative criteria as is proposed in the "NbClust" package [30]. Thus, in this study the optimal number of clusters was determined using the "NbClust" package in R-statistics with the default settings [30]. This algorithm combines 30 indices that estimate the optimal number of clusters and combines the results of all 30 into a majority vote. All indices have advantages as disadvantages and the majority vote eliminates the possibility of large effect due to sometimes arbitrary choices for one particular index.

Supplemental results

NbClust agreement

Among the 23 indices that were used to classify the correct number of clusters within the dataset (7 indices did not result in an optimal number of clusters):

- 7 proposed 2 as the best number of clusters
- 2 proposed 3 as the best number of clusters
- 2 proposed 4 as the best number of clusters
- 2 proposed 5 as the best number of clusters
- 1 proposed 6 as the best number of clusters
- 4 proposed 7 as the best number of clusters
- 2 proposed 9 as the best number of clusters
- 1 proposed 12 as the best number of clusters
- 2 proposed 15 as the best number of clusters

The optimal number of clusters was also 2 based on the Hubert index and on D-index.

Supplemental tables

Table S1: Rationale for selected biomarkers.

Marker	Rationale
ANG1/2	Cytokine involved in the control of microvascular permeability. The ratio of ANG1/2 indicates an increased permeability.
Antithrombin	Anti-coagulant effect
D-dimer	Fibrin degradation product; marker of activation of coagulatory response.
E-Selectin	Endothelial activation in a pro-inflammatory environment
Fractalkine	Chemokine, most strong associated with the recruitment of monocytes and T-cells.
GM-CSF	Pro-inflammatory cytokine
ICAM-1	Leukocyte migration
IFN- γ	Pro-inflammatory response, more associated with viral than bacterial resolution.
IL-10	Anti-inflammatory (mostly Th1) response in a hyperinflammatory environment
IL-13	Immune regulation and induction of MMPs in the airways
IL-1β	Pro-inflammatory cytokine
IL-6	Pro-inflammatory cytokine
IL-8	Chemokine, most strongly associated with the recruitment of neutrophils.
MMP-8	Marker of neutrophil activation; breaks down extracellular matrix.
P-Selectin	Endothelial activation in a pro-inflammatory environment
PAI-1	Inhibitor of fibrinolysis
TIMP1	Inhibitor of MMPs.
TNF-a	Pro-inflammatory cytokine
tPA	Effector for fibrinolysis

List of markers: Interleukin (IL)–1β, IL–6, IL–8, tumor necrosis factor alpha (TNF–α), IL–10, IL–13, interferon gamma (IFN–γ), granulocyte macrophage–colony stimulating factor (GM–CSF), soluble E–Selectin, soluble P–Selectin, fractalkine, plasminogen activator inhibitor (PAI)–1, D–dimer, tissue plasminogen activator (tPA), antithrombin, soluble intercellular adhesion modecule-1 (ICAM-1), matrix metalloproteinase–8 (MMP8), tissue inhibitor of of metalloprotease 1 (TIMP1), angiopoetin (ANG)1 and ANG2

Table S2: Missing data that needed imputation.

Variable	Number / percentage missing	
ICU mortality	0	0%
Phenotype	0	0%
APACHE IV	1	<1%
Gender	0	0%
APPS	147	21%
Berlin definition	0	0%
Pulmonary cause for ARDS	0	0%

Table S3: Regression coefficient for biomarkers concentrations.

Variable	Coefficient	SE	Wald	P-value
Intercept	-22.3	2.7	-8.1	<0.001
IL-6	3.98	0.48	8.4	<0.001
IFN-y	2.04	0.30	6.8	<0.001
ANG1/2	2.28	0.36	6.4	<0.001
PAI-1	2.21	0.37	6.0	<0.001

Biomarker concentrations were put into the model as 10 log transformer concentrations in pmol/L. SE: standard error.

Table S4: Phenotypes versus clinical characteristics in validation cohort.

		Uninflamed phenotype N=118		Reactive phenotype N=128		P
Age		63	(52-72)	60	(50-68)	0.11
Male		68	(57.6)	87	(68)	0.12
APACHE IV Score		73	(57-97)	90	(72-120)	<0.001
APACHE IV Acute Physiology Score		61	(47-83.5)	79	(58-104)	<0.001
Admission type	**Medical**	94	(79.7)	94	(73.4)	
	Elective surgery	10	(8.5)	11	(8.6)	
	Emergency surgery	14	(11.9)	23	(18)	
Chronic renal insufficiency		13	(11)	16	(12.5)	0.85
Chronic respiratory insufficiency		16	(13.6)	9	(7)	0.10
COPD		11	(9.3)	14	(10.9)	0.85
Diabetes mellitus		24	(20.3)	17	(13.3)	0.16
Immune deficiency		19	(16.1)	25	(19.5)	0.48
Current drinking status (alcohol)		7	(5.9)	12	(9.4)	0.35
Systemic corticosteroids (before ICU)		12	(10.2)	15	(11.7)	0.83
Direct hit for ARDS		75	(63.6)	75	(58.6)	0.45
Berlin definition	**Mild**	56	(47.5)	53	(41.4)	0.65
	Moderate	45	(38.1)	55	(43)	
	Severe	17	(14.4)	20	(15.6)	
Maximal inspiratory pressure		21	(15-25)	24	(18-31)	0.021
PaO2/FiO2		208	(170-280)	168.9	(134-215)	<0.001
PEEP		8	(5-12)	10	(8-12)	0.006
Tidal volume/kg predicted body weight		7.1	(6.3-8.5)	7.3	(6.5-8.8)	0.289
APPS		5	(4-5.2)	5	(4-6)	0.035
SOFA: Circulation		3	(1-4)	4	(1-4)	<0.001
SOFA: CNS		0	(0-1)	0	(0-1)	0.85
SOFA: Coagulation		0	(0-0)	0	(0-2)	<0.001
SOFA: Liver		0	(0-0)	0	(0-0)	0.002
SOFA: Renal		0	(0-0)	0	(0-0)	0.002
SOFA: Respiratory		0	(0-1)	1	(0-3)	<0.001
SOFA: Total score		3	(2-3)	3	(3-4)	0.02
Days on mechanical ventilation		6.5	(2-12)	6	(3-15)	0.327
ICU length of stay		8	(5-14)	8	(4-17)	0.85
Days free of MV at day 28		21	(14-26)	9.5	(0-23)	<0.001
ICU Mortality		16	(13.6)	48	(37.5)	<0.001
30-Day Mortality		26	(22)	50	(39.1)	0.008

Data is presented as the median with inter-quartile range for continuous variables and as number with percentage for categorical variables. The P-value is calculated by the Kruskall-Wallis test for continuous variables and by Fisher's exact for categorical variables. Definitions for the variables are given in the definition table at the end of the manuscript

Table S5: Excluded patients

		Excluded (N=173)		Included (N=700)		P
Age		62.0	(50-72)	62.0	(51-71)	0.589
Male		99.0	(57.2)	443.0	(63.3)	0.15
APACHE IV Score		82.0	(65-94.5)	82.0	(63-106.5)	0.407
APACHE IV Acute Physiology Score		67.0	(54-82.5)	70.0	(52-92)	0.363
Admission type	**Medical**	130.0	(75.1)	498.0	(71.1)	0.03
	Elective surgery	18.0	(10.4)	88.0	(12.6)	
	Emergency surgery	23.0	(13.3)	114.0	(16.3)	
Berlin definition	**Mild**	75.0	(43.4)	262.0	(37.4)	0.071
	Moderate	65.0	(37.6)	331.0	(47.3)	
	Severe	33.0	(19.1)	107.0	(15.3)	
SOFA: Total score		7.0	(5-10)	8.0	(6-11)	0.003
ICU Mortality		42.0	(24.3)	184.0	(26.3)	0.64
30-Day Mortality		47.0	(27.2)	212.0	(30.3)	0.510

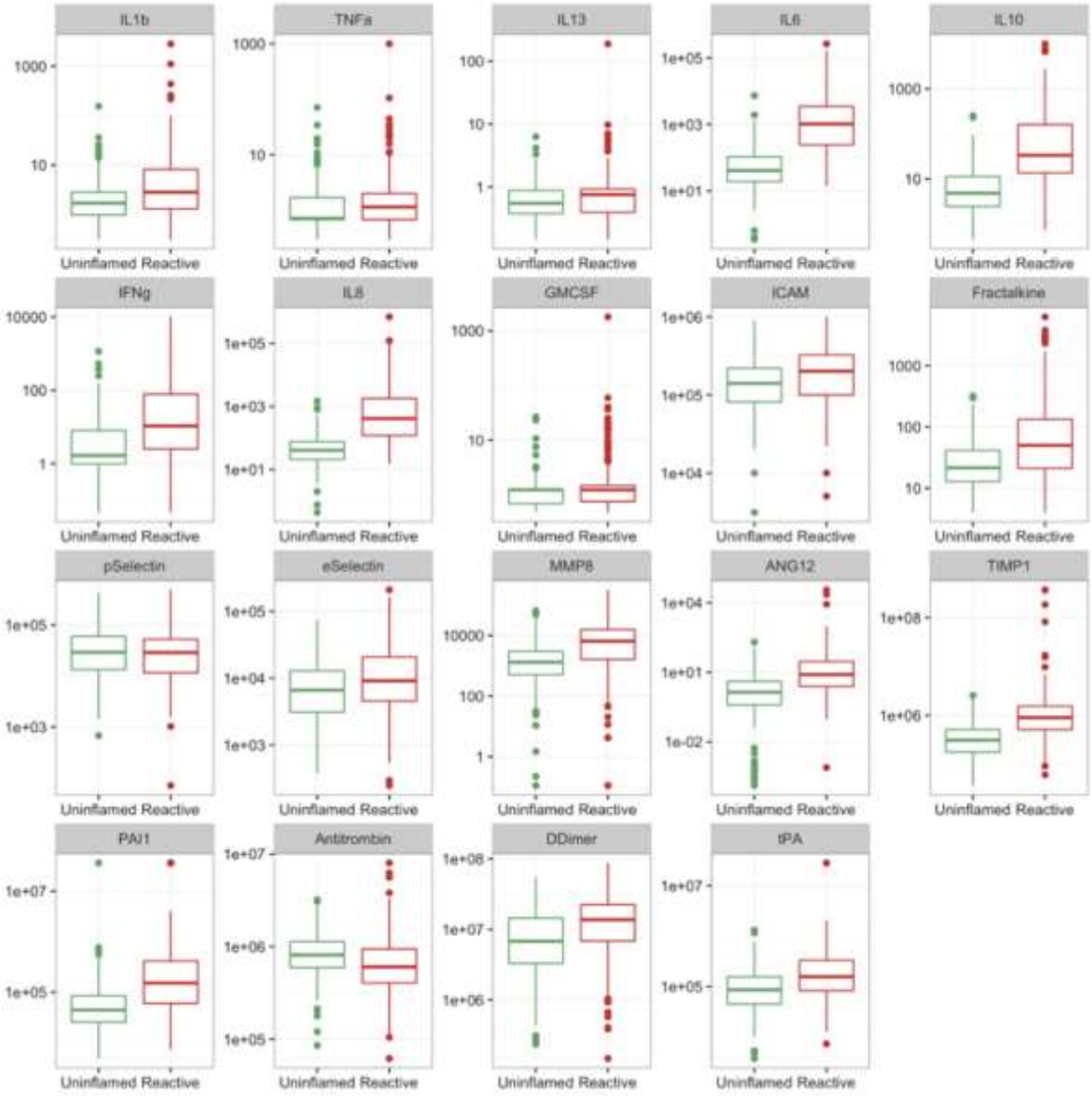
Table S6: Range of biomarker concentrations

Marker	Median	25%	75%	Min	Max
ANG1/ANG2	3	1	13	0	36061
Antitrombin	751306	501454	1068225	61900	8006500
D-Dimer	10466991	4452825	19408249	147452	87563000
eSelectin	8414	3914	18470	147	292899
Fractalkine	33	14	79	LLOQ	6329
GMCSF	LLOQ	LLOQ	LLOQ	LLOQ	1792
ICAM	170894	95968	288974	3134	1000000
IFNg	7	LLOQ	23	LLOQ	10000
IL10	13	5	48	LLOQ	10000
IL13	LLOQ	LLOQ	LLOQ	LLOQ	188
IL1b	3	LLOQ	5	LLOQ	2828
IL6	164	41	1126	LLOQ	265390
IL8	109	44	466	LLOQ	693916
MMP8	2294	860	7830	LLOQ	451953
PAI1	77853	32539	214381	2366	35911000
pSelectin	29705	13326	59364	70	495831
TIMP1	526767	269128	993155	36989	367990000
TNFa	LLOQ	LLOQ	2	LLOQ	1002
tPA	121913	61160	258671	3749	28180000

Biomarker concentration in pmol/L, expect for ANG1/2, which is the ratio of the two markers and thus has no unit. LLOQ is below the lower the lowest limit of quantification.
List of markers: Interleukin (IL)–1β, IL–6, IL–8, tumor necrosis factor alpha (TNF–α), IL–10, IL–13, interferon gamma (IFN–γ), granulocyte macrophage–colony stimulating factor (GM–CSF), soluble E–Selectin, soluble P–Selectin, fractalkine, plasminogen activator inhibitor (PAI)–1, D–dimer, tissue plasminogen activator (tPA), antithrombin, soluble intercellular adhesion modecule-1 (ICAM-1), matrix metalloproteinase–8 (MMP8), tissue inhibitor of of metalloprotease 1 (TIMP1), angiopoetin (ANG)1 and ANG2

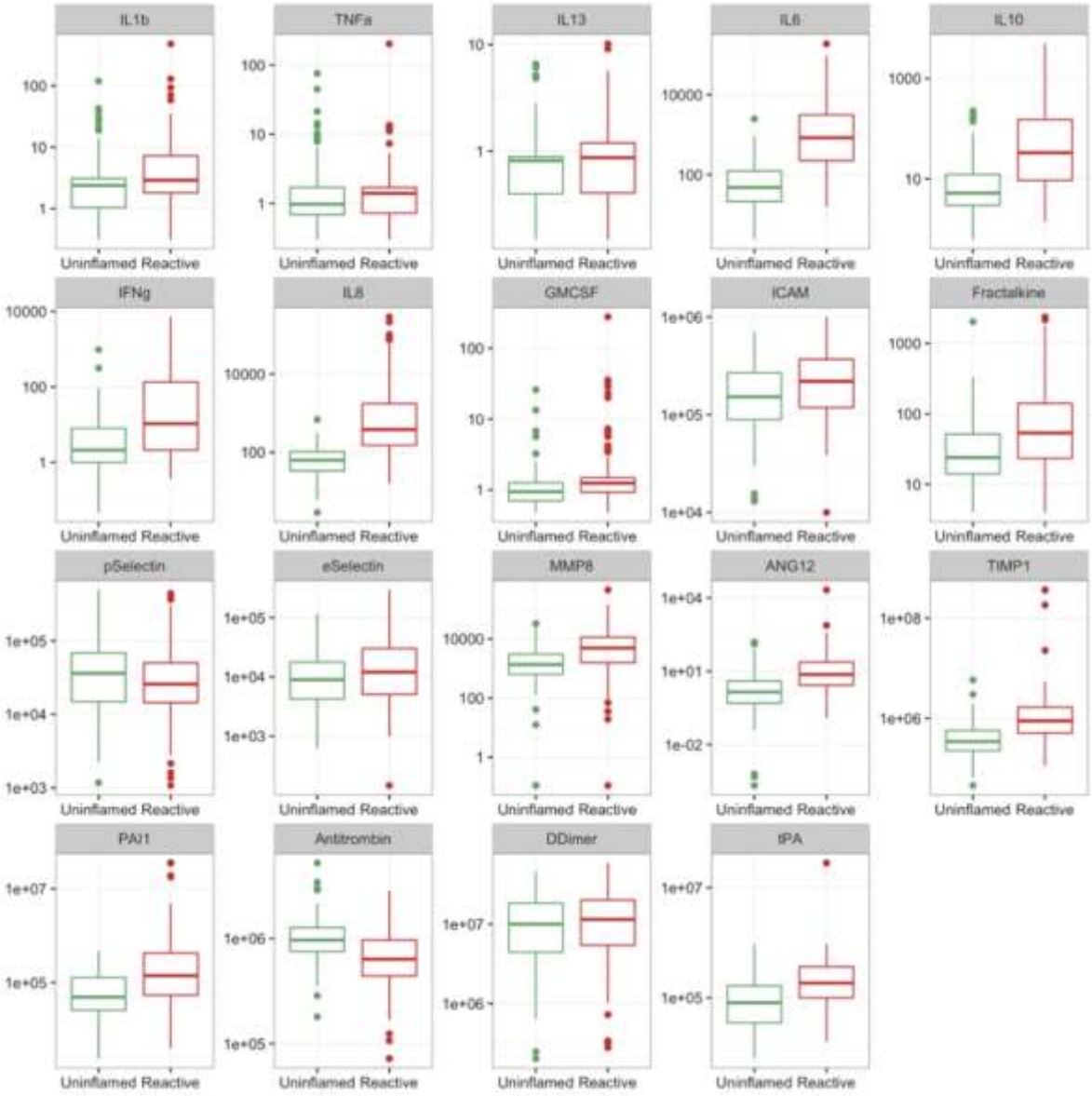
Online figures

Figure S1: Biomarker concentrations per biological phenotype in training cohort.



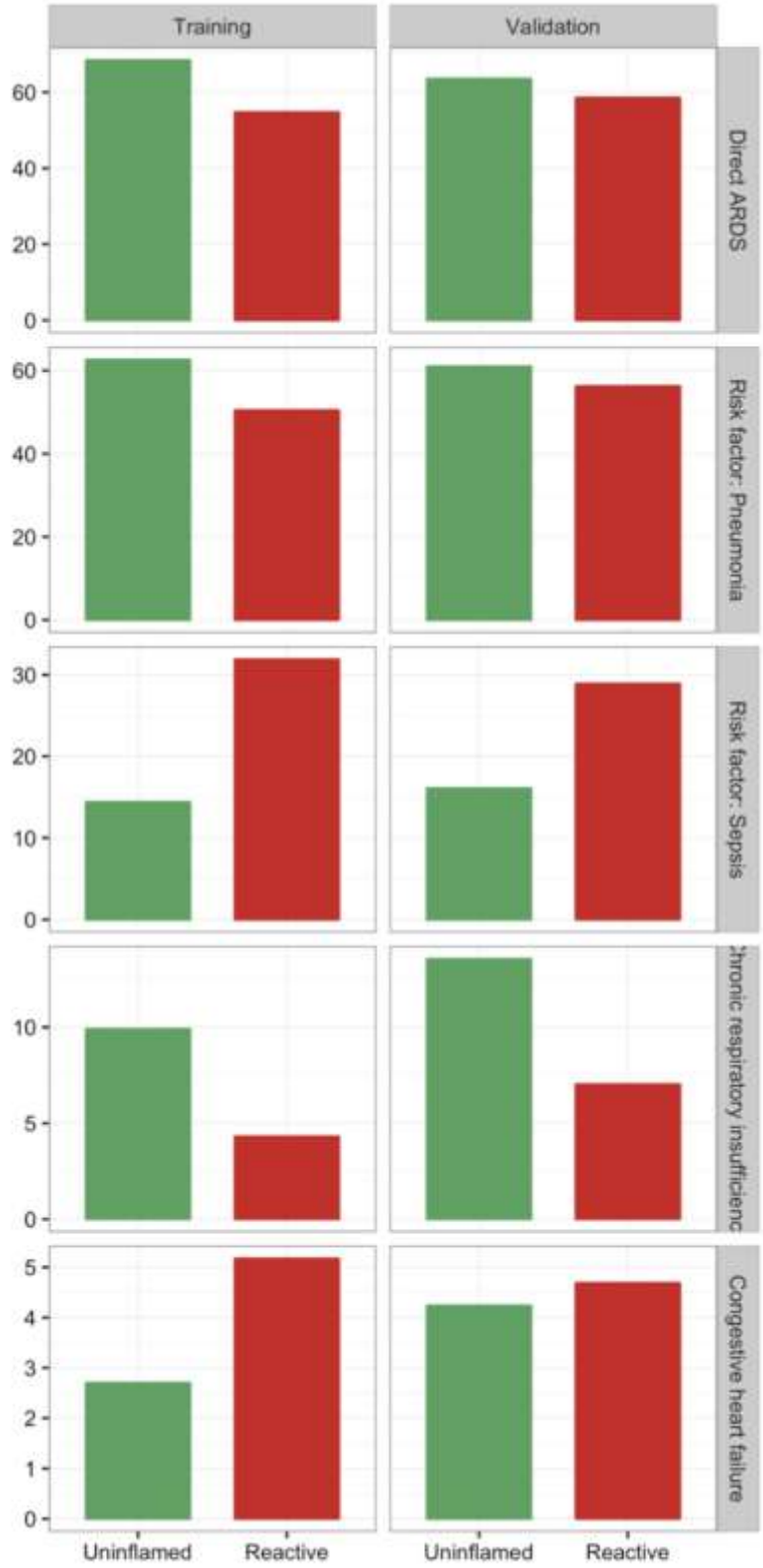
Legend: Y-axis = concentration of biomarker in pmol/L on a log10 scale. The limits of the Y-axis are different between the different plots. Green = ‘uninflamed’ phenotype. Red = ‘reactive’ phenotype.

Figure S2: Biomarker concentrations per biological phenotype in validation cohort.



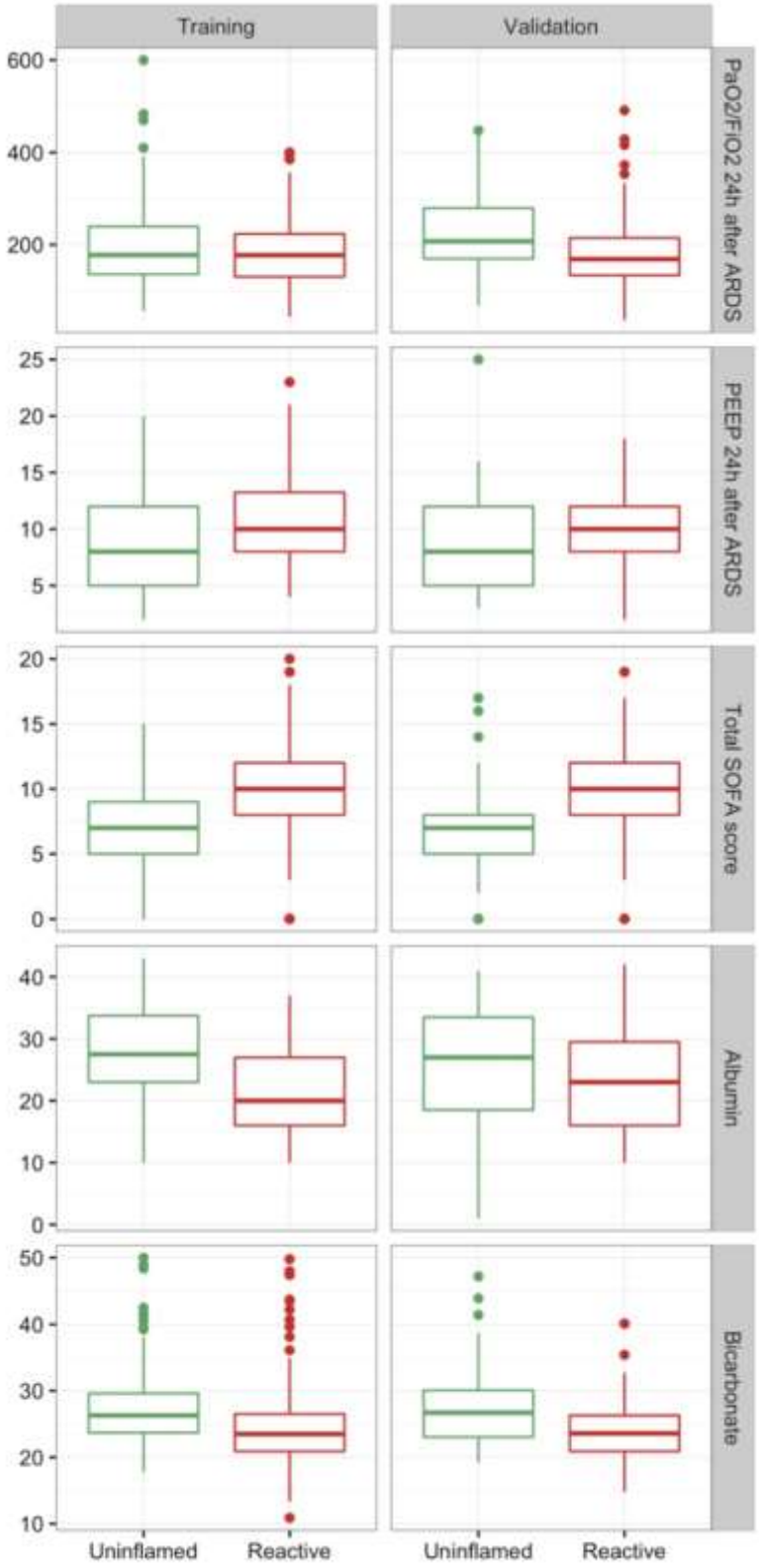
Legend: Y-axis = concentration of biomarker in pmol/L on a log10 scale. The limits of the Y-axis are different between the different plots. Green = ‘uninflamed’ phenotype. Red = ‘reactive’ phenotype.

Figure S3: Characteristics of phenotype; categorical variables.



Prevalence in percentage (Y-axis) of several categorical variables, stratified for training and validation cohort (P<0.05 for all). The limits of the Y-axis differ between the graphs.

Figure S4: Characteristics of phenotypes; continuous variables.



Median, interquartile range (box) and range (dots) for several continuous variables, stratified per training and validation cohort (P<0.05 for all). The limits of the Y-axis differ between the graphs.