Supplementary Material

Methods

Study Population and Sample Size

All patients in the study were admitted to the intensive care unit and were mechanically ventilated. Characteristics of the trial design and baseline population description for the individual trials are summarized in **Table S1**.

Sample size calculations are not feasible for model development; however, for validation, a minimum of 10 events per independent variable in multivariate regression modelling is recommended.[1] Another study recommended a minimum of 100 events and 100 non-events in the validation cohort in multivariate regression models containing six variables,[2] which was the maximum number of variables considered for parsimonious modeling in this study. Sample sizes in both validation cohorts, SAILS and HARP-2, met these recommendations.

Latent Class Analysis

Latent class analysis (LCA) is a form of mixture modeling that uses available data in a heterogeneous population to identify otherwise concealed subgroups. In the derivation dataset, the following biomarkers were used for LCA- interleukins 6 and 8 (IL-6, IL-8), plasminogen activator inhibitor-1 (PAI-1), protein C, soluble intercellular adhesion molecule-1 (sICAM-1), soluble tumor necrosis factor receptor-1 (sTNFr-1), surfactant protein D (SPD) and von Willebrand factor (vWF). A list of all the variables used during LCA modelling are presented in **Table S2**. All continuous variables underwent z-scale transformation. Additionally, non-normally distributed variables were log-transformed prior to z-scale transformation.

Four separate models consisting of one, two, three, and four classes were built. Once the model with the optimal number of classes was determined for the population, average probabilities for class assignment were generated for each observation. A priori, in line with our previous work, patients were assigned to a phenotype if the LCA-generated probability for that phenotype was 0.5 or greater. These class assignments served as the 'reference-standard' against which parsimonious model accuracy was tested.

Variable (Feature) Selection

LCA, like all finite mixture models, uses a combination of probability densities and regression analysis. Stepwise regression models to classify phenotype would lead to either over-fitting or, more likely, perfect fitting models, particularly given that the phenotypes in LCA were derived using the same variables that would be predictors in the regression analysis. To circumnavigate this issue of data circularity, two recursive partitioning machine learning algorithms, classification tree with Bootstrapped AGGregatING (bagging) and random forest, were used to identify the most important classifier variables in the derivation dataset. Least absolute shrinkage and selection operator (LASSO), was also used to identify influential classifier variables. The derivation dataset was randomly partitioned into 'training' (75%) and 'testing' (25%) datasets for the development and hyperparameter tuning of all machine learning models. Ten-fold cross-validation was used for tuning recursive partitioning models.

Classification Tree with Bagging

Classification tree analysis uses recursive partitioning algorithms to split a dataset using discrete cut-points of variables such as to best separate the data according to the dependent variable (phenotype). At each node all variables are interrogated to partition the data, and the best variable, and its optimal cut-off, are selected to split the node[3]. A complexity parameter (cp) is used to control the size of the tree by specifying and penalizing complexity at each split. A 10-fold cross-validation was used to determine the optimal cp.

Individual classification trees are subject to high variability. Minor perturbations in the training subset can lead to drastic changes in composition of the models. Bootstrap Aggregating (Bagging) is a simple ensemble procedure that reduces variance in model prediction by combining several classification trees[4]. Repeated bootstrap permutations of the training dataset were performed in order to generate 25 classifier trees. The aggregate classifier tree was used to

determine the most important variables. Classification tree analysis and bagging were analysed using R Studio with the 'rpart' (https://cran.r-project.org/web/packages/rpart) and 'ipred' (https://cran.r-project.org/package=ipred) packages respectively.

Random Forest

Random forest is similar to bagging in that it is an ensemble recursive partition technique where numerous trees are generated (>1000). The key difference between the two methods, aside from the number of trees, is that in random forest, for each tree a 'random' selection of predictors are used to grow the tree. The number of predictors used at each node are also selected randomly. Full trees are grown to the terminal node without pruning.[5] Individual trees may have poor predictive abilities, but the aggregate of these numerous trees greatly enhances accuracy, making random forest one of the most accurate machine learning algorithms. Numerous trees are generated using this method to form a forest. Classification is determined by a majority vote of all the trees in the forest. 'RandomForest' package in R Studio (https://CRAN.Rproject.org/package=randomForest) was used for Random Forest.

Least absolute shrinkage and selection operator (LASSO)

LASSO is a regression based technique that penalizes the absolute weight of coefficients such that all variable coefficients are shrunk towards zero.[6] Depending on the value of the coefficients and the tuning parameter (λ), the coefficients of minimally contributing predictor variables are set to zero, thereby eliminating them from the model. In this instance, minimized λ was sought using 10-fold cross validation. Sequential models were generated with increased values of λ starting at the minimal value. In order to satisfy our objective of generating a parsimonious model, we chose optimal λ as the smallest value that gave us a model with fewer than 8 variables. These retained variables were taken forward for consideration as contributors in the final nested LR model. All non-normal continuous variables were log-transformed prior to analysis. LASSO analysis was performed in R studio using 'glmnet' package (https://cran.r-project.org/package=glmnet).

Selecting the Most Important Variables

For the classification trees, at a given node, the primary or surrogate tree-splitting variable is attributed a goodness to split score, depending on how well the variable splits the population. The sum of goodness to split scores at each node of the tree was used to generate a variable importance score for all variables in the model. A score of 100 was assigned to the variable with highest score. All other variables were scaled against this score. These scores were used to determine the most important variables in the classification tree model. For the random forest models, we used the mean decrease Gini to determine the most important variables. The Gini impurity index is a measure of impurity at each node.[5] The index score for a homogeneous population would be zero and it reaches maximum value for a heterogeneous population. The total decrease in node impurity as a consequence of splitting with a variable and averaged over the entire tree gives the mean decrease Gini.

Missing Data

Despite the data being derived from randomized controlled trials, there were several variables with missing data. For model generation using classification tree, random forest, and LASSO regression, cases with missing values for biomarkers were excluded from the study. For missing categorical variables, data were complete for sex, race, and ARDS risk factor. In the derivation dataset, there were 141 missing values for vasopressor use at time of randomization. Data on vasopressor-use on the day of enrollment (Day 0) was used to complete these missing values. Phenotype derived using LCA from data with missing vasopressor values in the derivation dataset were highly concordant with phenotypes derived using the Day 0 data to complete missing values for vasopressor. The classes were the same using the two different datasets in 99.5% of the patients.

Missing values for continuous predictor variables (**Table S7**) were assumed to be missing-at-random and handled using multiple imputation by chained equation (MICE).[7 8] Plateau pressure was excluded from this part of the analysis due to high missingness (26%). Mean airway pressure was used as a surrogate for plateau pressure to determine the most important variables. In total we imputed 5 different data sets (n = 5). Analysis protocol for model generation and

variable selection using random forest and LASSO were repeated in each imputed model yielding the same six variables as the most important (data not presented).

Evidenced by our previous work, it was highly likely that biomarkers would constitute many of the most influential variables in predicting class. Given that the main objective of the study was feature selection rather than model parameter estimation, omitting observations with missing biomarker data seemed the most likely strategy to yield informative output. However, in order to ascertain that bias was not introduced secondary to missingness as a consequence of using complete cases, we repeated all the analyses using data where all missing values, including biomarkers, were imputed using identical techniques described above. In these analyses, the same six variables emerged as the most important (data are not presented). In addition, the characteristics of the study and missing cohorts were similar (**Table S8**). We used 'mice' package in R studio to impute the data (https://cran.r-project.org/web/packages/mice).

Development of Ancillary Classifier Models

The ancillary variables all consisted of a minimum of three and maximum of four variables. For each model, either interleukin-8 or interleukin-6 served as the base variable in the logistic regression model as both these variables were the most important variables by some margin compared to the other four most important variables (**Table S3** and **Figure S1**). Youden indices for the models were calculated in the derivation dataset, and along with a probability cut-off of 0.5, were used to assign phenotypes in the derivation dataset and in HARP-2. LCA-derived phenotypes were the reference standard against which model accuracy was tested.

Tables

Table S1. Selected cohort characteristics of the individual randomized control trials and baseline clinical characteristics of the patients included for model development and validation in the presented study

	ARMA	ALVEOLI	FACTT	SAILS	HARP-2	START
Patients in original trial (n)	861	549	1000	745	537	60
Patients in current study (n)	473	549	1000	715[a]	510[a]	58[a]
Study Years	1996-1999	1999-2003	2000-2005	2010-2013	2010-2014	2014-2017
Criteria used to define ARDS	AECC	AECC	AECC	Berlin	Berlin	Berlin
Time to enrollment from ARDS diagnosis	< 36 hours	< 36 hours	< 48 hours	< 48 hours	< 48 hours	< 96 hours
Intervention Studied	Low tidal volume	High PEEP	Conservative fluid strategy	Rosuvastatin	Simvastatin	Mesenchymal stromal cells
Treatment arms included in the study	Intervention only	Both	Both	Both	Both	Both
Tidal Volume/PBW (mL/KG)	10.1 (± 2.0)	8.1 (± 2.0)	7.4 (± 1.7)	6.7 (± 1.2)	8.1 (± 2.7)	6.2 (± 0.9)
PEEP (cm H_2O)	8 (5 – 10)	10 (5 – 12)	10 (5 – 12)	10 (5 – 11)	NA	10 (8 – 14)
PaO_2/FiO_2 (mmHg)	132 (± 60)	128 (± 58)	132 (± 63)	139 (± 64)	128 (± 55)	106 (± 40)
Plateau Pressure (cmH_2O)	29 (24 – 34)	26 (22 –31)	26 (22 – 30)	24 (19 – 28)	24 (20 – 28)	26 (22 – 30)
APACHE III score[b]	82 (± 29)	94 (± 32)	94 (± 31)	91 (± 28)	19 (± 7)[b]	100 (± 32)
Vasopressor at enrollment, n (%)	176 (37%)	144 (26%)	327 (33%)	407 (55%)	332 (65%)	41 (71%)
Ventilator free days, median (IQR)	14 (0 – 23)	18 (0 – 24)	17 (0 – 23)	20 (0 – 25)	13 (0 – 22)	6 (0 – 23)
Mortality at 90 days[c], n (%)	143 (30%)	148 (27%)	284 (28%)	204 (27%)	147 (29%)	19 (33%)[c]

APACHE = Acute Physiology, Age, Chronic Health Evaluation; NA = Not Available. a: Compared to the original trial cohort, fewer patients from these cohorts were analysed in this study due to lack of pertinent biomarker data; b : In the HARP-2 data, the APACHE II score is presented; c: In the START trial the mortality is at day 60. Vasopressor at enrollment was a yes / no dichotomous variable, Ventilator-free days was calculated to day 28. AECC = American-European Consensus Conference Criteria; PEEP = Positive end-expiratory pressure.

Table S2 Comparison of class defining variables between phenotypes in the *derivation* dataset

Class defining variables for initial LCA model	Hypo-inflammatory (n = 1431)	Hyper-inflammatory (n = 591)	P-value
Age (years)	50.6 (± 17)	50.1 (± 17)	0.53
Gender: Female	642 (45%)	259 (44%)	0.71[b]
Race: White	1033 (72%)	376 (64%)	0.0002[b]
Body mass index (BMI)	28.2 (± 7.2)	27.3 (± 7.4)	0.009
ARDS risk factor: Pneumonia	632 (44%)	205 (35%)	<0.0001[b]
ARDS risk factor: Sepsis	233 (16%)	245 (41%)	
ARDS risk factor: Other	566 (40%)	141 (24%)	
Temperature (°C)	38.4 (± 0.93)	38.6 (± 1.1)	<0.0001
Heart rate (beats.min^{-1})	120 (± 21)	137 (± 21)	<0.0001
Systolic blood pressure (mmHg)	92 (± 17)	79 (± 14)	<0.0001
Respiratory rate (breaths.min^{-1})	31 (25-39)	35 (29-40)	<0.0001[a]
Urine output (L over previous 24 hours)	2.3 (± 1.6)	1.9 (± 1.7)	<0.0001
Vasopressor use at baseline	259 (18%)	388 (66%)	<0.0001
PaO_2/FiO_2 ratio (mmHg)	135 (± 61)	119 (± 58)	<0.0001
$PaCO_2$ (mmHg)	40.3 (± 9.6)	36.1 (± 8.9)	<0.0001
Minute ventilation (L.min^{-1})	11.6 (± 3.5)	14.7 (± 4.5)	<0.0001
Tidal Volume (mL)	509 (± 138)	539 (± 138)	<0.0001
Plateau Pressure (cmH_2O)	26 (22-30)	30 (24-34)	<0.0001[a]
Positive end-expiratory pressure (cmH_2O)	8 (5-10)	10 (6-13)	<0.0001[a]
Hematocrit (%)	29.9 (± 6.0)	29.5 (± 6.6)	0.28
White cell count ($10^3/\mu L$)	15.1 (± 11.7)	13.7 (± 12.3)	0.02
Platelets ($10^3/\mu L$)	204 (± 127)	131 (± 102)	<0.0001
Sodium (mmol/L)	137 (± 5)	137 (± 6)	0.02
Glucose (mg/dL)	133 (± 56)	122 (± 69)	0.0007
Creatinine (mg/dL)	0.9 (0.7-1.4)	1.7 (1.1-2.7)	<0.0001[a]
Bicarbonate (mmol/L)	23.1 (± 4.9)	17.3 (± 5.0)	<0.0001
Albumin (g/dL)	2.3 (± 0.6)	2.0 (± 0.6)	<0.0001
Bilirubin (mg/dL)	1.4 (± 2.4)	2.4 (± 3.5)	<0.0001
Interleukin-6 (pg/mL)	116 (49-279)	933 (308-3026)	<0.0001[a]
Interleukin-8 (pg/mL)	23 (16-49)	133 (60-414)	<0.0001[a]
Soluble tumor-necrosis factor receptor-1 (pg/mL)	3225 (2236-5104)	7452 (4565-10879)	<0.0001[a]
Intercellular adhesion molecule 1 (ng/mL)	959 (589-1561)	1239 (742-2072)	<0.0001[a]
Protein C (% control)	96.0 (± 57)	53.5 (± 38)	<0.0001
Plasminogen activator inhibitor 1 (ng/mL)	56 (36-86)	107 (71-170)	<0.0001[a]
Surfactant Protein-D (ng/mL)*	125 (60-275)	86 (42-166)	<0.0001[a]
Von Willebrand Factor (% control)*	203 (112-343)	337 (199-538)	<0.0001[a]

P values represent the 2-sample t-test unless annotated (a = Mann-Whitney U test, b = chi-square test). *These variables were unavailable for latent class analysis in the validation dataset.

Table S3 Goodness to split score from the classification tree with bagging model of the top seven variables of importance in the derivation dataset.

Variable	Goodness to split
IL-6	100
IL-8	89.3
Vasopressors	82.6
sTNFR-1	75.8
Bicarbonate	75.8
Protein C	52.6
Creatinine	49.4

sTNFR-1 = Soluble tumour necrosis factor receptor-1

Table S4 Model composition with variable coefficients of ancillary parsimonious models and their area under the receiver operator characteristic curve (AUC), sensitivity and specificity used in the validation dataset. A. Probability cut-off set as Youden Index estimated from the derivation dataset to assign phenotypes B. Probability cut off set as 0.5 to assign phenotypes.

Model Coefficients								A. Youden Index		B. Probability ≥ 0.5	
Intercept	IL-8*	$HCO3^-$	sTNFR1*	vasopressor	Protein C*	IL-6*	AUC (95% CI)	Sensitivity	Specificity	Sensitivity	Specificity
-10.6110	1.2902	-0.2326	1.0732	--	--	--	0.95 (0.93 – 0.96)	0.93	0.83	0.79	0.94
-13.1351	1.3947	-0.2145	1.1818	2.1398	--	--	0.96 (0.95 – 0.97)	0.95	0.81	0.90	0.89
-18.4764	1.3013	--	1.3367	2.3439	--	--	0.94 (0.92 – 0.95)	0.97	0.66	0.89	0.83
4.7119	--	-0.2707	--	--	-1.3522	0.9402	0.90 (0.88 – 0.92)	0.82	0.83	0.73	0.90
4.0323	--	-0.2581	--	1.7412	-1.3805	0.9191	0.92 (0.90 – 0.94)	0.93	0.70	0.81	0.87
-11.9593	--	-0.2436	1.1903	--	--	1.0138	0.94 (0.92 – 0.95)	0.93	0.74	0.82	0.89
-13.7440	--	-0.2257	1.2785	1.8375	--	1.0089	0.95 (0.93 – 0.96)	0.96	0.72	0.90	0.84
-19.4464			1.4711	2.0793		0.9107	0.92 (0.91 – 0.94)	0.97	0.58	0.89	0.74

IL8 = Interleukin-8, IL6 = Interleukin-6, sTNFR1 = soluble tumour necrosis factor receptor-1. * Coefficients were derived using the logarithmic ($\log_e$) transformed values of the variables. For IL-8, sTNFR1, IL-6, and protein C, a value of 1 was added to the measured value prior to enable log transformation.

Table S5. 2 X 2 Table of phenotype assignment by latent class analysis and by parsimonious model (Interleukin-6, soluble tumour-necrosis factor receptor 1, vasopressor-use) in HARP-2. Table S5A Parsimonious model probability cut-off to assign hyper-inflammatory phenotype was the Youden Index (≥ 0.276). Table S5B Parsimonious model probability cut-off to assign hyper-inflammatory phenotype was ≥ 0.5.

A. Youden Index	**Parsimonious model derived Hypo-inflammatory class**	**Parsimonious model derived Hyper-inflammatory Class**	**Total**
LCA Assigned Hypo-inflammatory class	234 (specificity 0.62)	99	333
LCA Assigned Hyper-inflammatory Class	17	160 (sensitivity 0.93)	177
Total	251	259	

B. ≥ 0.5	**Parsimonious model derived Hypo-inflammatory class**	**Parsimonious model derived Hyper-inflammatory Class**	**Total**
LCA Assigned Hypo-inflammatory class	292 (specificity 0.77)	41	333
LCA Assigned Hyper-inflammatory Class	38	139 (sensitivity 0.88)	177
Total	330	180	

Table S6. Phenotype assignment in the START trial. Phenotype assignment using the three best parsimonious models using the probability cut-offs as either the Youden Index for the models extracted from the derivation dataset or a value greater than 0.5. Day 60 mortality for each subphenotype are also presented. Using chi-squared test, mortality was significantly higher in the hyper-inflammatory phenotype compared to the hypo-inflammatory phenotype ($p < 0.002$ for all models except where annotated; a: $p = 0.011$, b: $p = 0.023$).

	3-variable Model > 0.5 cut-off		3-variable Model Youden Index cut-off		4-variable Model > 0.5 cut-off		4-variable Model Youden Index cut-off	
Phenotypes	**Count**	**60-Day Mortality**	**Count**	**60-Day Mortality**	**Count**	**60-Day Mortality**	**Count**	**60-Day Mortality**
Hypo-Inflammatory	43 (74%)	8 (19%)	41 (71%)	8 (20%)	39 (67%)	8 (21%)[a]	35 (60%)	7 (20%)[b]
Hyper-Inflammatory	15 (26%)	11 (73%)	17 (29%)	11 (65%)	19 (33%)	11 (58%)[a]	23 (40%)	12 (52%)[b]

Table S7. Summary of the missing data in the merged LCA dataset (n=2022) that was used for latent class analysis and the complete cases dataset used as the 'derivation dataset' for model development. Multiple imputation using chained equations was used for missing data in the complete biomarker cases column.

	Merged all cases (n =2022) (used for LCA)	Complete biomarker cases (n=1558) (Derivation dataset)
Body mass index	160	118
Age	0	0
Temperature	3	2
Systolic Blood Pressure	4	3
Heart rate	3	2
PaO_2/FiO_2 ratio	1	0
Tidal Volume	309	232
Minute Ventilation	36	28
PEEP	7	6
Mean airway pressure	279	186
$PaCO_2$	98	76
Respiratory rate	3	2
Urine Output	77	63
Haematocrit	9	7
WBC	43	26
Platelets	18	16
Sodium	13	4
Creatinine	34	23
Glucose	34	22
Albumin	317	219
Bilirubin	346	273
Bicarbonate	20	17
Protein C	262	-
PAI-1	189	-
Interleukin-6	224	-
Interleukin-8	217	-
SolubleTNFR-1	298	-
ICAM-1	262	-
Surfactant Protein D	351	-
Von Willebrand Factor	288	-

(PEEP = Positive end expiratory pressure, PAI-1 = Plasminogen activation factor 1, TNFR-1 = Tumour necrosis factor receptor 1, ICAM-1 = Intercellular adhesion molecule 1)

Table S8. Comparison of key characteristics between the studied population for variable selection and the cohort excluded due to missing data in the derivation dataset.

	Study Cohort	Missing Cohort	P-Value
Number (%)	1558 (77%)	464 (23%)	NA
Hyper-inflammatory Phenotype	452 (29%)	139 (30%)	0.74*
Source Study:			
ARMA	268 (17%)	205 (44%)	< 0.0001*
ALVEOLI	500 (32%)	49 (11%)	
FACTT	790 (51%)	210 (45%)	
Mean PaO_2/FiO_2 (mmHg)	131 (± 61)	128 (± 60)	0.34✝
Median Ventilator Free Days (IQR)	17 (0-23)	15 (0-23)	0.16▲
Mortality at day 90	437 (28%)	138 (30%)	0.52*

P-values are representative of: * Chi-square test, ✝ 2-sample t-test, ▲ Wilcoxon test.

FIGURES

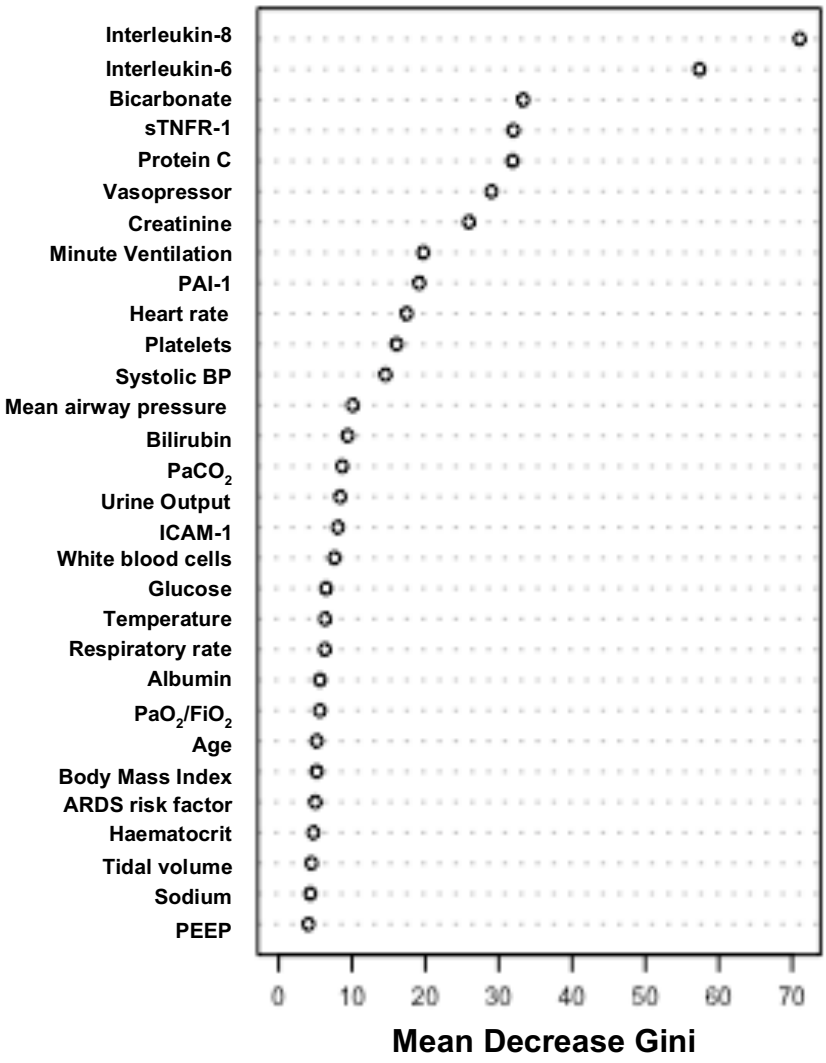

Figure S1 Mean decrease of Gini score and accuracy in the random forest classifier model to select the most important variables in the model. sTNFR1 = soluble tumor necrosis factor receptor 1, Vasopressor = vasopressor use, VE = minute ventilation, PAI-1 = plasminogen activation factor-1, HR = heart rate, SBP = systolic blood pressure, $PaCO_2$ = Arterial pressure of carbon dioxide, MAP = mean airway pressure, UO = urine output over last 24 hours, WBC = white blood cell count, ICAM-1 = Intra-cellular adhesion molecule-1, RF = risk factor, VT = tidal volume, RR = respiratory rate, HCT = hematocrit, PEEP = Positive end-expiratory pressure, BMI = body mass index, PF ratio = PaO_2/FiO_2 ratio.

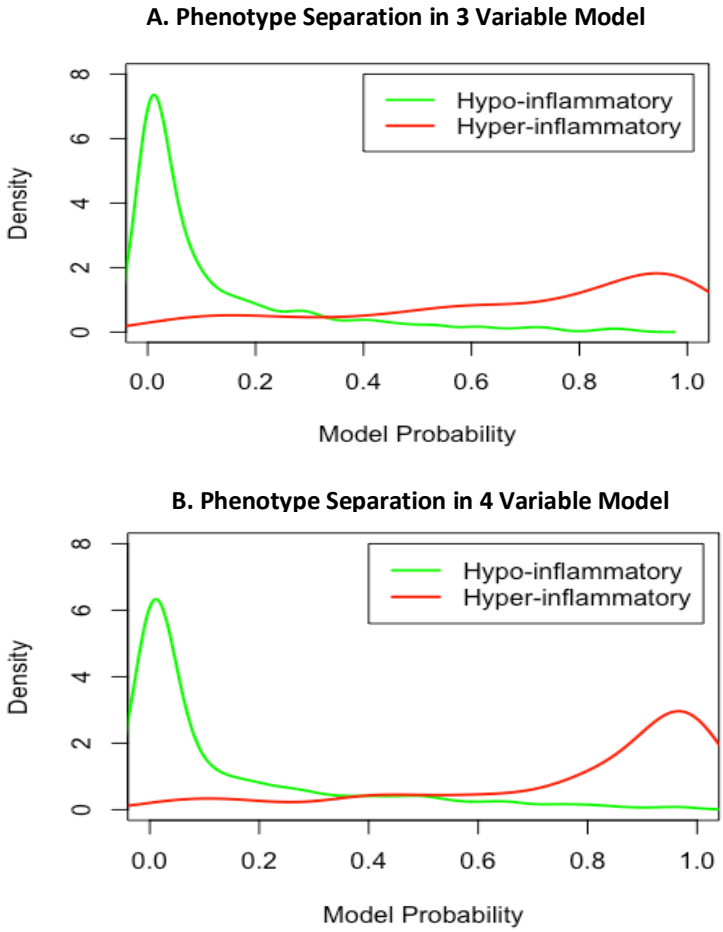

Figure S2 Kernel density function plot for the distribution of ARDS phenotypes as defined by latent class analysis (LCA) plotted against the probability of belonging to the hyper-inflammatory subphenotype as generated by the parsimonious regression models in the Validation cohort. A. 3-variable model (interleukin-8, bicarbonate, and protein C); B. 4-Variable Model (interleukin-8, bicarbonate, Protein C, and Vasopressor use).

Reference:

1. Peduzzi P, Concato J, Feinstein AR, Holford TR. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. *J Clin Epidemiol* 1995; **48**(12): 1503-10.
2. Vergouwe Y, Steyerberg EW, Eijkemans MJ, Habbema JD. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *J Clin Epidemiol* 2005; **58**(5): 475-83.
3. Breiman L. Classification and regression trees. Belmont, Calif.: Wadsworth; 1984.
4. Breiman L. Bagging predictors. *Machine Learning* 1996; **24**(2): 123-40.
5. Breiman L. Random forests. *Machine Learning* 2001; **45**(1): 5-32.
6. Tibshirani R. The lasso method for variable selection in the Cox model. *Stat Med* 1997; **16**(4): 385-95.
7. Van Buuren S G-OK. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software* 2011; **45**(3): 1-67.
8. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? *Int J Methods Psychiatr Res* 2011; **20**(1): 40-9.