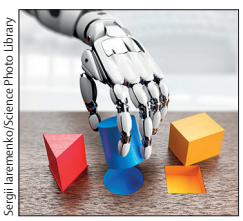
# A critical approach to personalised medicine in ARDS


Sergii Iaremenko/Science Photo Library

Published Online
January 13, 2020
https://doi.org/10.1016/
S2213-2600(19)30352-2

See Articles page 247

Acute respiratory distress syndrome (ARDS) is characterised by different pathogenetic pathways leading to similar clinical presentations. The mechanisms leading to ARDS are now better understood than ever; nevertheless, mortality remains high, probably due to biological heterogeneity, which hinders identification of an effective therapy.[1] Therefore, recent studies have focused on identifying specific ARDS phenotypes in an attempt to improve diagnosis, optimise therapeutic interventions, and allow better selection of patients for future randomised controlled trials.[2] Two different ARDS phenotypes (hyperinflammatory and hypoinflammatory) have been identified, showing differences in response to therapy and outcome.[3] The hyperinflammatory phenotype is characterised by increased plasma levels of inflammatory biomarkers, vasopressor dependence, and prevalence of sepsis, as well as lower serum bicarbonate. However, this classification was based on analysis of 37 clinical parameters, which would not be feasible—or would at least be extraordinarily challenging—at the bedside. Within this context, a simple method to identify ARDS phenotypes would be a step towards the development of individually targeted therapeutic strategies. Similarly, in patients undergoing open abdominal surgery, two different phenotypes based on pre-operative plasma inflammatory serum biomarker concentrations have been identified and were able to distinguish patient groups with different incidences of postoperative pulmonary complications.[4]

In *The Lancet Respiratory Medicine*, Pratik Sinha and colleagues[5] reported the results of a retrospective analysis of data pooled from cohorts of large ARDS randomised controlled trials, aimed at developing a simple model to facilitate phenotypic identification in patients with ARDS at the bedside. In an elegant, hypothesis-driven study with sophisticated statistical analyses, the authors tested the prognostic validity of their models in two external ARDS clinical trial datasets (START[6] and HARP-2[7]). The two ARDS phenotypes, hyperinflammatory and hypoinflammatory, could be accurately identified with a simple logistic regression model using three or four variables (interleukin-8, bicarbonate, and protein C, with the optional addition of vasopressor use). Machine learning systems were used to optimise these analyses.

Some limitations of this model preclude its immediate use in the clinical setting. First, the analysis included highly selected patients with ARDS from different randomised controlled trials done at different times, which no longer reflect current clinical practice. Second, ARDS is a syndrome caused by several aetiologies, and different causative insults affect outcomes in different ways. ARDS diagnosis is based on a combination of clinical and radiological criteria; it includes neither cellular and humoral components of the inflammatory response nor molecular and genetic biomarkers. Even if routine testing of molecular biomarkers were available widely and performed routinely, there is no clear threshold for any such markers to distinguish ARDS from ARDS-like syndromes, which limits these markers' utility in clinical practice. Furthermore, standardisation of the period of time for data collection has an important effect on identification of patient phenotypes. Finally, none of the variables included in the models are specific for ARDS, and real-time testing of plasma biomarkers is currently unavailable in many intensive care units.

The foundational assumptions of personalised medicine are twofold: first, the patient's phenotype should be correctly identified; second, treatment needs to be individually targeted and effective for that specific phenotype. In diseases with a single factor, such as cancer, a personalised approach seems to be associated with better outcomes. In ARDS, however, multiple factors contribute to disease progression. Thus, even if individual phenotypes are identified correctly, this might not ensure the efficacy of specific treatment. In this context, recent evidence suggests that individualised mechanical ventilation strategies might not be effective to improve outcomes both in perioperative medicine[8] and in ARDS.[9] Care must be taken to refrain from prematurely positive interpretations of reanalyses of previously collected data, which have not been validated in prospective patient cohorts nor in randomised controlled trials. To date, no guidelines that include different ARDS phenotypes have been published. For personalised medicine to be applied to the clinical care of ARDS, clear experimental and clinical action plans must be developed. First, the development and implementation of clinical models, including both physiological and biological variables, that can easily

identify the different phenotypes of patients with ARDS at the bedside requires organisational structure and investment. Second, these models must be validated in prospective observational studies recruiting large samples of patients worldwide. Third, the validated models should then be tested in randomised controlled trials, hypothesising that different treatments might differently affect outcomes according to a pre-planned phenotypic stratification. In conclusion, the idea of applying personalised medicine to ARDS management on the basis of specific phenotypes is attractive, but it must still be clinically proven before translation to the clinical setting.

We declare no competing interests.

*Lorenzo Ball, Pedro L Silva, Patricia R M Rocco, *Paolo Pelosi*
**ppelosi@hotmail.com**

Department of Surgical Sciences and Integrated Diagnostics, University of Genoa, Genoa, Italy (LB, PP); Anaesthesia and Intensive Care, IRCCS San Martino Policlinico Hospital, 16131 Genoa, Italy (LB, PP); and Laboratory of Pulmonary Investigation, Carlos Chagas Filho Institute of Biophysics, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil (PLS, PRMR)

1 Matthay MA, Zemans RL, Zimmerman GA, et al. Acute respiratory distress syndrome. *Nat Rev Dis Prim* 2019; **5:** 18.
2 Hendrickson CM, Calfee CS. A new frontier in ARDS trials: phenotyping before randomisation. *Lancet Respir Med* 2019; **7:** 830–31.
3 Calfee CS, Delucchi K, Parsons PE, et al. Subphenotypes in acute respiratory distress syndrome: latent class analysis of data from two randomised controlled trials. *Lancet Respir Med* 2014; **2:** 611–20.
4 Serpa Neto A, Bos LD, Campos PPZA, et al. Association between pre-operative biological phenotypes and postoperative pulmonary complications: an unbiased cluster analysis. *Eur J Anaesthesiol* 2018; **35**: 702–09.
5 Sinha P, Delucchi KL, McAuley DF, O'Kane CM, Matthay MA, Calfee CS. Development and validation of parsimonious algorithms to classify acute respiratory distress syndrome phenotypes: a secondary analysis of randomised controlled trials *Lancet Respir Med* 2020; published online Jan 13. https://doi.org/10.1016/S2213-2600(19)30369-8.
6 Matthay MA, Calfee CS, Zhuo H, et al. Treatment with allogeneic mesenchymal stromal cells for moderate to severe acute respiratory distress syndrome (START study): a randomised phase 2a safety trial. *Lancet Respir Med* 2019; **7:** 154–62.
7 McAuley DF, Laffey JG, O'Kane CM, et al. Simvastatin in the acute respiratory distress syndrome. *N Engl J Med* 2014; **371**: 1695–703.
8 Ferrando C, Soro M, Unzueta C, et al. Individualised perioperative open-lung approach versus standard protective ventilation in abdominal surgery (iPROVE): a randomised controlled trial. *Lancet Respir Med* 2018; **6**: 193–203.
9 Constantin J-M, Jabaudon M, Lefrant J-Y, et al. Personalised mechanical ventilation tailored to lung morphology versus low positive end-expiratory pressure for patients with acute respiratory distress syndrome in France (the LIVE study): a multicentre, single-blind, randomised controlled trial. *Lancet Respir Med* 2019; **7:** 870–80.

# *FLT1*: a potential therapeutic target in sepsis-associated ARDS?

Published **Online**
January 23, 2020
https://doi.org/10.1016/S2213-2600(19)30404-7

See **Articles** page 258

In the past 10 years, genome-wide association studies (GWAS) have become a standard approach to improve our understanding of the aetiology of complex diseases. Many such studies have been done, and multiple associations reported for a wide range of diseases or subphenotypes. However, for most diseases, the biggest challenge remains to be addressed—moving from showing association with a genetic marker (typically a single-nucleotide polymorphism [SNP] chosen because it is informative about genetic linkage in a region of the human genome) to gaining an understanding of the functional mechanisms underlying the association.

In *The Lancet Respiratory Medicine*, Beatriz Guillen-Guio and colleagues[1] report a GWAS of sepsis-associated acute respiratory distress syndrome (ARDS). They make the case that a variant at the Fms-related tyrosine kinase 1 (*FLT1*) gene locus, which encodes vascular endothelial growth factor receptor 1 (VEGFR-1), is protective for risk of developing the disease. The discovery stage was done using 274 cases (patients with sepsis-associated ARDS) and 316 controls (patients with sepsis without ARDS) from the GEN-SEP cohort. The MESSI and SepNet cohorts were used for replication. Although no variant reached conventional genome-wide significant association in discovery, rs9508032 at the *FLT1* locus was genome-wide significant after meta-analysing results across the three populations (odds ratio 0·61, 95% CI 0·41–0·91; $p=5{\cdot}2\times10^{-8}$). In a further look up, some evidence was also seen for association between SNPs at the vascular endothelial growth factor A (*VEGFA*) locus (VEGF-A being the ligand for VEGFR-1) and disease risk, although these associations were well below the level required to be genome-wide significant. As GWAS studies go, the number of cases in this study is

Kateryna Kon/Science Photo Library

# Development and validation of parsimonious algorithms to classify acute respiratory distress syndrome phenotypes: a secondary analysis of randomised controlled trials

*Pratik Sinha, Kevin L Delucchi, Daniel F McAuley, Cecilia M O'Kane, Michael A Matthay, Carolyn S Calfee*

## Summary

**Background** Using latent class analysis (LCA) in five randomised controlled trial (RCT) cohorts, two distinct phenotypes of acute respiratory distress syndrome (ARDS) have been identified: hypoinflammatory and hyperinflammatory. The phenotypes are associated with differential outcomes and treatment response. The objective of this study was to develop parsimonious models for phenotype identification that could be accurate and feasible to use in the clinical setting.

**Methods** In this retrospective study, three RCT cohorts from the National Lung, Heart, and Blood Institute ARDS Network (ARMA, ALVEOLI, and FACTT) were used as the derivation dataset (n=2022), from which the machine learning and logistic regression classifer models were derived, and a fourth (SAILS; n=715) from the same network was used as the validation test set. LCA-derived phenotypes in all of these cohorts served as the reference standard. Machine-learning algorithms (random forest, bootstrapped aggregating, and least absolute shrinkage and selection operator) were used to select a maximum of six important classifier variables, which were then used to develop nested logistic regression models. Only cases with complete biomarker data in the derivation dataset were used for variable selection. The best logistic regression models based on parsimony and predictive accuracy were then evaluated in the validation test set. Finally, the models' prognostic validity was tested in two external ARDS clinical trial datasets (START and HARP-2) by assessing mortality at days 28, 60, and 90 and ventilator-free days to day 28.

**Findings** The six most important classifier variables were interleukin (IL)-8, IL-6, protein C, soluble tumour necrosis factor receptor 1, bicarbonate, and vasopressor use. From the nested models, three-variable (IL-8, bicarbonate, and protein C) and four-variable (3-variable plus vasopressor use) models were adjudicated to be the best performing. In the validation test set, both models showed good accuracy (AUC 0·94 [95% CI 0·92–0·95] for the three-variable model and 0·95 [95% CI 0·93–0·96] for the four-variable model) against LCA classifications. As with LCA-derived phenotypes, the hyperinflammatory phenotype as identified by the classifier model was associated with higher mortality at day 90 (87 [39%] of 223 patients *vs* 112 [23%] of 492 patients; p<0·0001) and fewer ventilator-free days (median 14 days [IQR 0–22] *vs* 22 days [0–25]; p<0·0001). In the external validation datasets, three-variable models developed in the derivation dataset identified two phenotypes with distinct clinical features and outcomes consistent with previous findings, including differential survival with simvastatin versus placebo in HARP-2 (p=0·023 for survival at 28 days).

**Interpretation** ARDS phenotypes can be accurately identified with parsimonious classifier models using three or four variables. Pending the development of real-time testing for key biomarkers and prospective validation, these models could facilitate identification of ARDS phenotypes to enable their application in clinical trials and practice.

**Funding** National Institutes of Health.



*Lancet Respir Med* 2020; 8: 247–57

Published Online January 13, 2020 https://doi.org/10.1016/S2213-2600(19)30369-8

See Comment page 218

**Department of Medicine, Division of Pulmonary, Critical Care, Allergy and Sleep Medicine** (P Sinha PhD, Prof M A Matthay MD, Prof C S Calfee MD), **Department of Anesthesia** (P Sinha, Prof M A Matthay, Prof C S Calfee), **and Department of Psychiatry** (Prof K L Delucchi PhD), **University of California, San Francisco, San Francisco, CA, USA**; **The Wellcome-Wolfson Institute for Experimental Medicine, Queen's University, Belfast, UK** (Prof D F McAuley MD, Prof C M O'Kane PhD); **and Regional Intensive Care Unit, The Royal Hospitals, Belfast, UK** (Prof D F McAuley)

Correspondence to: Dr Pratik Sinha, University of California San Francisco Medical Center, San Francisco, CA 94143–0111, USA **pratik.sinha@ucsf.edu**

## Introduction

Despite more than 50 years of research into acute respiratory distress syndrome (ARDS), disappointingly few clinical trials have resulted in positive findings. The few trials that have succeeded include low tidal volume ventilation, prone positioning, and fluid-conservative strategies.[1–3] Tellingly, all these interventions were designed to improve supportive care. No clinical trials testing pharmacological interventions in ARDS have identified a benefit. The broad clinical definition of ARDS, with the ensuing heterogeneity in cause and pathophysiology it captures, are increasingly seen as one of the reasons for these negative trials.[4]

To address the issue of heterogeneity, researchers have recently used latent class analysis (LCA) in ARDS. LCA is a form of mixture modelling that uses available data to identify unmeasured or latent subgroups in a heterogeneous population. Two phenotypes, termed hyperinflammatory and hypoinflammatory, have been consistently identified in five randomised controlled trial (RCT) cohorts of ARDS.[5–8] Mortality and other clinical outcomes are worse in the hyperinflammatory phenotype.

## Research in context

### Evidence before this study

Using latent class analysis (LCA), previous studies have consistently identified two phenotypes—hyperinflammatory and hypoinflammatory—across five randomised controlled trial (RCT) cohorts of acute respiratory distress syndrome (ARDS). These phenotypes have distinct biological and clinical characteristics with divergent clinical outcomes and differential responses to therapy in secondary analyses of RCTs. The complexity of the LCA models that identify the phenotypes is a major impediment to their application in the clinical setting. Whether parsimonious models using a selection of key variables could be used to identify the two ARDS phenotypes remains unknown. No formal literature search was done for this study.

### Added value of this study

Using an array of machine learning algorithms, the presented study identifies parsimonious models comprised of three or four variables that can accurately classify ARDS phenotypes in two external validation cohorts. The phenotypes identified using these parsimonious models shared similar characteristics and outcomes to phenotypes identified using LCA. The survival benefit observed with simvastatin in a previous analysis was also observed in the hyperinflammatory phenotype identified using the parsimonious model. In a recent trial testing the efficacy of mesenchymal stem cells in ARDS, the hyperinflammatory phenotype identified by parsimonious models was associated with significantly higher mortality at day 60.

### Implications of all the available evidence

Heterogeneity in ARDS is increasingly being recognised as a potential contributing factor to failed clinical trials. LCA-identified phenotypes offer researchers more biologically and clinically uniform subgroups to test hypotheses and interventions. With the simpler models described in this study, identification of the phenotypes could become more feasible and might herald a new era of prospective, phenotype-specific trials in ARDS.

Although some studies have used clinical data to identify phenotypes in ARDS that might be useful for prognostic enrichment,[9,10] LCA-derived ARDS phenotypes also offer the potential for predictive enrichment. In secondary analyses of two RCTs, the LCA-derived phenotypes responded differently to positive-end expiratory pressure[5] and fluid therapy.[6]

Recently, in a secondary analysis of the completed HARP-2 trial,[11] which assessed hydroxymethylglutaryl-CoA reductase inhibition with simvastatin in acute lung injury to reduce pulmonary dysfunction, a survival benefit was observed in the hyperinflammatory phenotype in patients randomly assigned to simvastatin compared with placebo.[7] No treatment effects were observed in the original RCT. These findings suggest a potential route for prognostic and predictive enrichment in ARDS trials. Although these results are promising, key barriers limit the identification of these ARDS phenotypes in clinical practice. Most notably, the complexity of the described LCA models, which can consist of up to 40 predictor variables, renders them impractical for prospective clinical use. The main hypothesis of this study was that a simpler model consisting of a maximum of six variables could accurately classify ARDS phenotypes.

In a previous study, a three-variable model was shown to identify these phenotypes with good accuracy.[6] The model, however, had several limitations. Most pertinently, the model used *z*-scaled values of the classifier variables, rendering them unsuitable for prospective use because prior knowledge of the variables' population distribution would be necessary. Additionally, the model was derived using a single RCT cohort and was variably accurate in independent cohorts, suggesting suboptimal stability.[6] Furthermore, differential treatment effects observed with the original LCA-models were not observed when patients were classified using this model. The primary objective of this study was to develop and validate parsimonious models that could ultimately be used prospectively to identify ARDS phenotypes.

## Methods

### Study design

In this retrospective modelling study, development and validation of the models proceeded in three steps. First, we developed the models using a combined cohort of three RCTs to improve model performance and increase their generalisability. Next, with LCA-derived phenotypes as the gold standard, we evaluated model performance of the two best-fitting models using a contemporaneous ARDS RCT from the same research network and two RCT datasets external to the network. Finally, we tested whether, as with LCA-derived phenotypes, a differential treatment effect with simvastatin was observed in phenotypes determined by these parsimonious models in one of the external RCTs.

### Datasets

We generated two datasets: a derivation dataset, from which machine learning and logistic regression classifier models were derived for variable selection and phenotype derivation, and an out-of-sample validation test set, used to evaluate the accuracy of two of the best-performing models. The derivation dataset was generated by combining three RCTs from the National Heart, Lung, and Blood Institute (NHLBI) ARDS Network: ARMA (high *vs* low tidal volume),[1] ALVEOLI

(high *vs* low positive end-expiratory pressure),[12] and FACCT (conservative *vs* liberal fluid management).[3] The validation dataset was derived from the most contemporaneous RCT from the same network, SAILS,[13] testing rosuvastatin versus placebo in sepsis-related ARDS. Selected trial and population baseline characteristics are presented in the appendix (p 5). Additional details on trial protocols and study populations can be found in the original studies.

For external validation, we used data from two recently completed RCTs that were not part of the NHBLI ARDS Network: HARP-2 and START. HARP-2 tested the efficacy of simvastatin (80 mg once daily) versus placebo in ARDS.[11] START was a phase 2a trial that tested the safety of intravenous human bone marrow-derived mesenchymal stromal cells for moderate to severe ARDS.[14] Details of study protocols and patient populations can be found in the original studies.[11,14]

### Primary analysis overview

An overview of the primary analysis plan is outlined in figure 1. Briefly, LCA was conducted on the derivation dataset, and the resultant phenotypes served as both the dependent variable for machine learning models that were developed for the nested regression variable selection and as the reference standard to test model performance. For the purposes of variable selection, the derivation dataset was split into a training dataset (75%) and a holdout dataset for hyperparameter tuning (25%). A ten-fold cross-validation was used for tuning the recursive partitioning algorithms. The most important variables were, in turn, used to develop nested logistic regression classifier models. Of these, the two best models were used for out-of-sample testing in the validation test set. LCA-derived phenotype assignment in the validation test set was generated in a previous study and served as the reference standard to test model accuracy.[8]

### LCA

We used LCA in the derivation dataset to identify the optimal number of classes that best fit the population. In line with our previous work, we used a combination of demographic, clinical, standard laboratory, and protein biomarkers, all from at or before the time of randomisation, as class-defining variables in the models (appendix p 6).[5,6] No clinical outcome variables or severity scores were used in the modelling procedures. Four separate models consisting of one, two, three, and four classes were built. Optimal model selection for the population was judged using the Bayesian information criterion (BIC), the Vuong-Lo-Mendell-Rubin (VLMR) likelihood ratio test, the number of observations in the smallest class (classes containing relatively small numbers were not considered clinically meaningful), and entropy. Further details on LCA procedures can be found in the appendix (p 2).

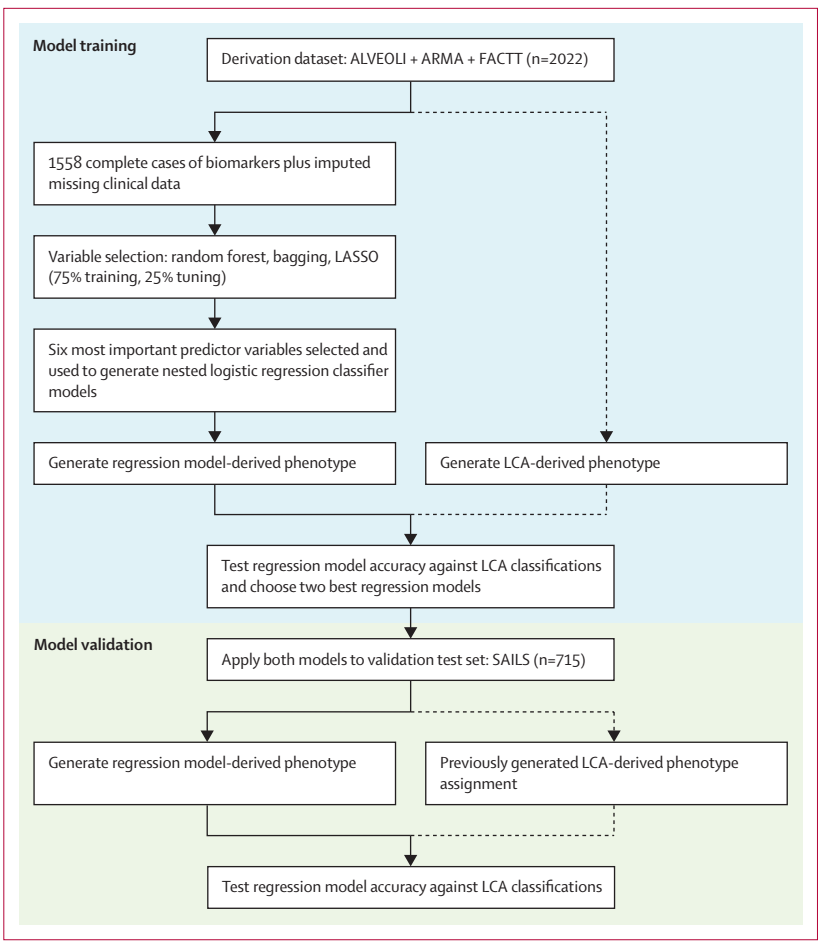



***Figure 1:*** **Overview of the primary analysis plan**
LCA=latent class analysis. Bagging=bootstrapped aggregating. LASSO=least absolute shrinkage and selection operator.

### Variable selection

See Online for appendix

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. For variable selection, both techniques are known to penalise categorical variables, particularly those with the fewest categories.[15] Therefore, a third method, least absolute shrinkage and selection operator (LASSO), was also used to identify important classifier variables. To limit the complexity of our models, we decided a priori to limit the maximum number of variables to six for the final modelling, based on previous experience.

Previous work indicated that protein biomarkers were likely to be essential components of regression classifier models.[6] Therefore, only cases with complete biomarker data in the derivation dataset were used for variable selection (figure 1). Multiple imputation with chained equations was used to impute missing clinical data in the derivation dataset (appendix pp 3–4). To select the most

| | Interleukin-8 | Bicarbonate | Protein C | Vasopressor use | Soluble TNF receptor 1 | Interleukin-6 | AUC (95% CI) | AIC |
|---|---|---|---|---|---|---|---|---|
| Model 1 | Yes | No | No | No | No | No | 0·86 (0·84–0·88) | 1268 |
| Model 2 | Yes | Yes | No | No | No | No | 0·92 (0·90–0·93) | 1005 |
| Model 3 | Yes | Yes | Yes | No | No | No | 0·95 (0·93–0·96) | 835 |
| Model 4 | Yes | Yes | Yes | Yes | No | No | 0·96 (0·95–0·97) | 719 |
| Model 5 | Yes | Yes | Yes | Yes | Yes | No | 0·97 (0·96–0·98) | 638 |
| Model 6 | Yes | Yes | Yes | Yes | Yes | Yes | 0·97 (0·97–0·98) | 585 |

Table shows the six most important variables included in a sequential nested regression model. The order in which the variables entered the model was determined by the findings of a stepwise regression analysis with all variables in the model. TNF=tumour necrosis factor. AUC=area under the receiver operator characteristic curve. AIC=Akaike information criterion.

***Table 1:*** **Nested model composition and accuracy in the derivation dataset**

important variables, a goodness-to-split score was used for the bagging model and the Gini impurity index for the random forest model (appendix pp 2–3). For the LASSO modeling, the tuning parameter λ was sequentially altered such that there were fewer than eight variables in the final model. The six most important classifier variables common to all three machine learning algorithms were then used to generate nested logistic regression models.

### Logistic regression models

The top six variables identified by the machine learning models were used in a forward stepwise regression using the derivation dataset. Nested logistic regression models of increasing complexity were generated by sequential addition of the variables. The order in which variables were entered into the nested models was determined by findings of stepwise regression analysis.

Model performance was assessed by generating receiver operating characteristic (ROC) curves and calculating the area under the ROC curve (AUC) and 95% CIs for each model. The Akaike information criterion (AIC) and the Youden Index were also generated for each model. Likelihood ratio tests were used to compare nested model performance. In both datasets, data that were not normally distributed were log-transformed for regression modelling. To test for interaction between outcome and predictor variables, the analysis was repeated by introducing first-order interaction terms to the models.

### Model performance in the validation datasets

We decided a priori to take forwards two nested logistic regression models and their coefficients from the derivation dataset to test in the validation dataset. The two best models were determined by a combination of accuracy in the derivation dataset and model parsimony. These primary classifier models were used to generate probabilities for phenotype assignment in the validation test set. For each model, the hyperinflammatory phenotype was assigned using a probability cutoff of 0·5 or the Youden Index generated in the derivation dataset, accepting probabilities of those or higher for phenotype classification. These phenotype assignments were used to calculate sensitivity, specificity, and accuracy of the models. DeLong's test was used to compare ROC curves and the $\chi^2$ test was used to compare model performance. As a sensitivity analysis, the accuracy of ancillary models using permutations of the six best predictor variables were also tested in the validation test set, with each model composed of three or four variables.

To test the validity of the models in identifying phenotypes in non-NHLBI ARDS Network RCTs, model performance was evaluated in external validation datasets: START (primary classifier) and HARP-2 (ancillary), using either the primary or ancillary classifier models depending on data availability. In both studies, clinical and biological data at enrolment were used to assign phenotype using the classifier model developed as above, and outcome data (mortality and ventilator-free days) were used to assess the prognostic validity of phenotype classification. For mortality, the timepoints used were the same as the ones used in the original trials (days 28 and 90 for the NHLBI ARDS Network trials and HARP-2; days 28 and 60 for the START trial). Ventilator-free days were censored at day 28.

Assay procedures for plasma biomarker quantification can be found in the original studies.[5,7,8] In HARP-2, phenotypes identified by the ancillary classifier model applicable to the dataset were evaluated against prior LCA assignment.[7] Additionally, randomisation data in HARP-2 were used to evaluate treatment interaction with the primary classifier model-derived phenotypes and simvastatin. In START, LCA was not done due to insufficient sample size. The characterisation and appropriateness of the identified phenotypes were evaluated using clinical data and outcomes.

### Statistical analysis

Spearman's correlation coefficients were calculated to index agreement of class probabilities generated by LCA and the primary classifier models. Between-group differences were tested using the two-sample *t* test and Mann-Whitney *U* test depending on the distribution of the variable. Difference in outcome in phenotypes was tested using Pearson's $\chi^2$ test. For testing differential response to treatment by class for survival (time to death),

time-to-event Kaplan-Meier curves were compared using Wilcoxon test. LCA was conducted using Mplus software version 8.2. All other analyses were done using R Studio version 3.3.0.

### Role of the funding source

The funders of this study had no role in study design, data collection, data analysis, interpretation of the data, or writing of report. The corresponding author had full access to all the data and final responsibility to submit for publication. CSC and KLD also had access to all the raw data. CMO'K, DFM, and MAM had access to part of the raw data.

## Results

The derivation dataset was comprised of 2022 patients. In LCA, a two-class model best fit this dataset, showing significantly improved fit compared with the one-class model (p<0·0001). Further improvement in model fit was not observed when going to a three-class (p=0·35) or four-class (p=0·13) model. Good class separation was observed in the two-class model (entropy=0·83). There were 1431 patients (70·8%) classified as the hypoinflammatory and 591 (29·2%) as the hyperinflammatory phenotype. Mean probabilities for class membership were 0·96 (SD 0·1) for the hypoinflammatory class and 0·93 (0·1) for the hyperinflammatory class.

In the derivation dataset, the hyperinflammatory phenotype was associated with increased mortality at day 90 (267 [45%] *vs* 308 [22%]; p<0·0001) and with fewer ventilatory-free days (median 3 days [IQR 0–19] *vs* 20 days [1–24]; p<0·0001). Key characteristic differences between phenotypes are summarised in the appendix (p 6) and are in keeping with previous studies.Of the 2022 patients in the derivation dataset, 1558 (77%) had complete biomarker data and were used for the development of the primary classifier models.

The most important classifier variables from the bagging and random forest models are presented in the appendix (pp 7, 13). Using LASSO with λ=0·1, the seven predictor variables included in the final model were bicarbonate, interleukin (IL)-6, IL-8, plasminogen-activator inhibitor-1 (PAI-1), protein C, soluble tumour necrosis factor (TNF) receptor 1, and vasopressor use. Bicarbonate, IL-6, IL-8, protein C, soluble TNF receptor 1, and vasopressor use were common to all three machine-learning models and were therefore selected as the six best classifier variables for the parsimonious models.

Forward stepwise regression did not eliminate any of the six variables (table 1). No significant interactions were observed when first-order interaction terms were introduced in the models. Increasing model complexity with sequential addition of predictors led to significantly improved model performance (p<0·0001). There was, however, a relative plateauing of AUC and AIC in the four-variable, five-variable, and six-variable models

| | Derivation dataset (1558 complete cases) | Validation test set (715 complete cases) | p value* |
|---|---|---|---|
| Sex | .. | .. | 0·78† |
| Female | 709 (45·5%) | 363 (50·8%) | .. |
| Male | 849 (54·5%) | 352 (49·2%) | .. |
| Race | .. | .. | 0·0001† |
| White | 1107 (71·1%) | 564 (78·9%) | .. |
| Non-white | 451 (28·9%) | 151 (21·1%) | .. |
| Body-mass index, kg/m² | 28·1 (7·3) | 30·7 (10·1) | <0·0001 |
| Age (years) | 50·3 (16·4) | 54·1 (16·3) | <0·0001 |
| Temperature (°C) | 38·4 (1·0) | 38·1 (1·0) | <0·0001 |
| Systolic blood pressure (mmHg) | 89 (17) | 85 (16) | <0·0001 |
| Heart rate (bpm) | 125 (22) | 119 (23) | <0·0001 |
| $PaO_2/FiO_2$ ratio | 131 (61) | 139 (64) | 0·0070 |
| Tidal volume (mL) | 507 (132) | 414 (88) | <0·0001 |
| Minute ventilation (mL/min) | 12·5 (4·0) | 10·8 (3·2) | <0·0001 |
| PEEP (cm $H_2O$) | 10 (5–12) | 10 (5–10) | 0·90‡ |
| $PaCO_2$ (mmHg) | 39 (10) | 40 (11) | 0·011 |
| Respiratory rate (breath/min$^{-1}$) | 33 (26–40) | 32 (27–38) | 0·58‡ |
| Urine output (L/24 h) | 2·2 (1·7) | 1·6 (1·2) | <0·0001 |
| Haematocrit (%) | 30 (6) | 30 (6) | 0·73 |
| White blood cells (10³/μL) | 14·4 (11·1) | 15·7 (12·5) | 0·018 |
| Platelets (10³/μL) | 182 (121) | 185 (121) | 0·58 |
| Sodium (mmol/L) | 137 (6) | 138 (5) | 0·0058 |
| Creatinine (mg/dL) | 1·5 (1·4) | 1·5 (1·2) | 0·98 |
| Glucose (mg/dL) | 129 (61) | 125 (46) | 0·061 |
| Albumin (g/dL) | 2·2 (0·6) | 2·2 (0·6) | 0·41 |
| Bilirubin (mg/dL) | 1·6 (2·8) | 1·3 (1·8) | 0·0056 |
| Bicarbonate (mmol/L) | 21·3 (5·6) | 21·8 (5·6) | 0·077 |
| Protein C (% control) | 84·8 (54·3) | 80·4 (42·2) | 0·036 |
| PAI-1 (ng/mL)§ | 66 (40–110) | 4 (2–9) | <0·0001‡ |
| Interleukin-6 (pg/mL) | 179 (64–567) | 448 (174–1531) | <0·0001‡ |
| Interleukin-8 (pg/mL) | 33 (17–83) | 53 (25–137) | <0·0001‡ |
| Soluble TNF receptor 1 (pg/mL) | 3964 (2500–7176) | 5355 (3090–8876) | <0·0001‡ |
| ICAM-1 (ng/mL) | 1074 (632–1751) | 360 (236–511) | <0·0001‡ |
| APACHE III score | 91·2 (30·9) | 93·4 (28·2) | 0·10 |
| ARDS risk factors | .. | .. | <0·0001† |
| Trauma | 133 (8·5%) | 6 (0·8%) | .. |
| Sepsis | 369 (23·7%) | 140 (19·6%) | .. |
| Aspiration | 231 (14·8%) | 44 (6·2%) | .. |
| Pneumonia | 662 (42·5%) | 510 (71·3%) | .. |
| Other | 163 (10·5%) | 15 (2·1%) | .. |
| Vasopressor use on day of enrolment | 471 (30·2%) | 394 (55·1%) | <0·0001† |
| Ventilator-free days | 17 (0–23) | 20 (0–25) | 0·0009‡ |
| Mortality at 90 days | 437 (28·0%) | 199 (27·8%) | 0·96† |

$PaO_2/FiO_2$ ratio=ratio of partial pressure of oxygen in arterial blood to the fractional concentration of oxygen in inspired air. PEEP=positive-end expiratory pressure. $PaCO_2$=partial pressure of carbon dioxide in arterial blood. PAI-1=plasminogen activator inhibitor-1. TNF=tumour necrosis factor. ICAM-1=intercellular adhesion molecule-1. APACHE=Acute Physiology, Age, and Chronic Health Evaluation. ARDS=acute respiratory distress syndrome. *p values represent *t* test unless stated otherwise. †$\chi^2$ test. ‡Mann-Whitney *U* test. §Observed differences in values in PAI-1 might be due to different assays used for quantification.

***Table 2:*** **Comparison of the derivation dataset and validation test set in the primary analysis**

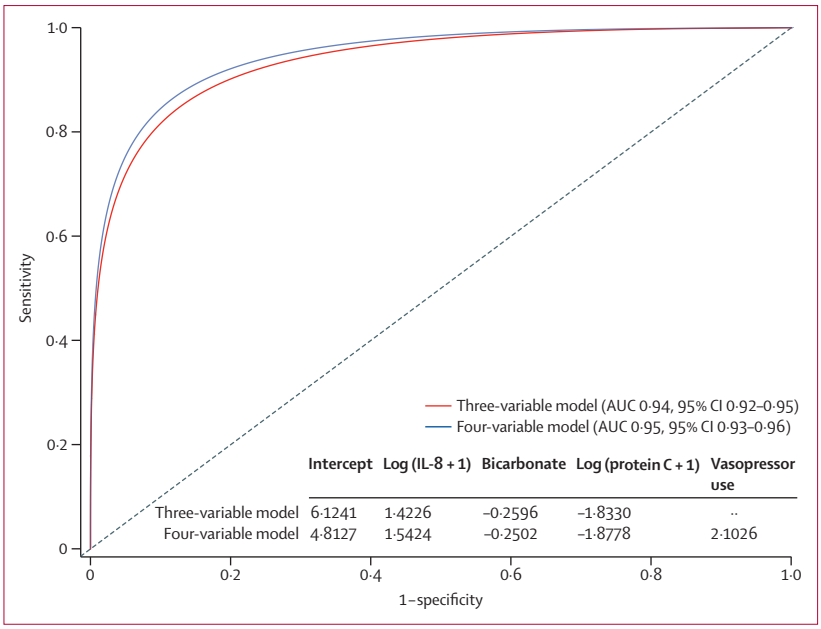



***Figure 2*: Receiver operating characteristic curves of the two best-performing regression models in the validation test set, with model coefficients**
AUC=area under the curve. IL-8=interleukin-8.

(table 1). The three-variable (IL-8, bicarbonate, and protein C) and four-variable (IL-8, bicarbonate, protein C, and vasopressor use) models were, therefore, considered best in terms of balancing classifying accuracy and model simplicity, and were thus defined as our primary classifier models. The Youden Index generated from the derivation dataset was 0·295 for the three-variable model and 0·301 for the four-variable model.

The performance of the models were next evaluated in the validation dataset. Differences in baseline between the derivation dataset and validation test set are summarised in table 2; of note, significant differences were observed in levels of biomarkers and vasopressor use between the cohorts. In the validation test set, the AUC was 0·94 (95% CI 0·92–0·95) for the three-variable model and 0·95 (0·93–0·96) for the four-variable model (figure 2). When setting the Youden Index as the probability cutoff to assign phenotype, the three-variable model had higher specificity than the four-variable model; however, the sensitivity of the four-variable model was higher (table 3). With the probability cutoff set at 0·5, specificity increased in both models to more than 0·9, with the three-variable model having higher specificity (table 3).

The median probability generated by the primary classifier models for belonging to the hyperinflammatory phenotype was 0·85 (IQR 0·68–0·97) for the three-variable model and 0·93 (0·79–0·99) for the four-variable model. The distribution of probabilities was sparse in the range of 0·3–0·7 (appendix p 14), suggesting good phenotype discriminatory properties of both models. The probabilities for phenotype assignment generated by the LCA model showed strong positive correlation with those generated by the primary classifier models ($r$=0·85 for the three-variable model, $r$=0·87 for the four-variable model; p<0·0001 for both).

For the three-variable model, the mean LCA-derived probability was lower for the misclassified patients compared with the correctly classified patients in both the hyperinflammatory (0·88 *vs* 0·98) and hypoinflammatory (0·89 *vs* 0·96) phenotype. This finding suggests that assignment of LCA-derived phenotypes was less certain in patients misclassified by the primary classifier models.

Compared with the hypoinflammatory phenotype, the hyperinflammatory phenotype was associated with higher mortality at day 90 (87 [39%] of 223 patients *vs* 112 [23%] of 492 patients; p<0·0001) and fewer ventilator free days (median 14 days [IQR 0–22] *vs* 22 days [0–25]; p<0·0001). These differences in clinical outcomes were consistent when the four-variable model was used to assign phenotype (data not shown) and when the Youden Index was used to assign class in both models (data not shown). Overall, differences in clinical outcomes were similar to the original LCA-derived phenotypes.

Details of the procedures for the ancillary classifier model development and testing can be found in the appendix (p 4). Most of the ancillary classifier models showed good accuracy in classifying phenotypes, with an AUC of at least 0·90 in the validation test set for all models (appendix p 8). Replacing the protein C variable with a soluble TNF receptor 1 variable resulted in similar AUCs in both the three-variable and four-variable models. Replacing IL-8 with IL-6 generally increased model sensitivity but reduced specificity.

When considering external validation of the models, validation of the primary classifier model in HARP-2 was not possible due to the unavailability of IL-8 and protein C; therefore, we used an ancillary three-variable model comprised of IL-6, soluble TNF receptor 1, and vasopressor use to classify phenotypes in this dataset (appendix p 8). 508 (94%) of 540 patients had complete data available to estimate classification probabilities. The AUC for the ancillary model was 0·92 (95% CI 0·89–0·94). Using the Youden Index from the derivation dataset for this ancillary model as the probability cutoff to assign the hyperinflammatory phenotype (≥0·276) resulted in a sensitivity of 0·93 and specificity of 0·62 (appendix p 9). Increasing the probability cutoff to at least 0·5 led to a sensitivity of 0·88 and specificity of 0·77. When a probability cutoff of 0·5 was used for phenotype assignment, 180 (35%) patients were classified as hyperinflammatory and 330 (65%) patients as hypoinflammatory. These proportions were similar to those for the original LCA-derived phenotypes (appendix p 9).[7]

Mortality at day 28 (66 [37%] *vs* 58 [18%]; p<0·0001) and at hospital discharge (73 [41%] *vs* 74 [22%]; p<0·0001) were significantly higher in the hyperinflammatory phenotype. The hyperinflammatory phenotype was also

| | Youden Index | | | Probability ≥0·5 | | |
|---|---|---|---|---|---|---|
| | Hyperinflammatory | Sensitivity | Specificity | Hyperinflammatory | Sensitivity | Specificity |
| Three-variable model: interleukin-8, bicarbonate, protein C | 40% | 0·84 | 0·87 | 31% | 0·74 | 0·95 |
| Four-variable model: interleukin-8, bicarbonate, protein C, vasopressors | 44% | 0·91 | 0·83 | 36% | 0·82 | 0·91 |

Table shows data for two probability cutoffs: Youden Index (three-variable model: 0·295; four-variable model: 0·301) and 0·5.

***Table 3:*** **Percentage of patients classified as hyperinflammatory phenotype, sensitivity, and specificity of the primary classifier models in the validation test set**

associated with fewer ventilator-free days (median 4 days [IQR 0–19] *vs* 17 days [0–23]; p<0·0001).

Significantly different survival curves were observed across patients stratified by the ancillary model-derived phenotype and treatment (figure 3; p<0·0001). In the hyperinflammatory phenotype, treatment with simvastatin was associated with significantly higher survival at 28 days compared with placebo (p=0·023). This pattern for survival was also similar at day 90, although the higher observed survival with simvastatin failed to reach statistical significance (overall p<0·0001; p=0·062 for simvastatin compared with placebo in the hyperinflammatory phenotype). These treatment effects were not observed in the hypoinflammatory phenotype. Overall, these findings were similar to our prior analysis using LCA-derived phenotypes.[7]

When externally validating the primary classifier models in START, biomarker data were available for 58 (97%) of the 60 patients. The three-variable and four-variable models both identified the phenotypes with similar prevalence as in our previous studies (60–74% in hypoinflammatory and 26–40% in hyperinflammatory; appendix p 10). Mortality at day 60 was significantly higher in the hyperinflammatory group regardless of the model or probability cutoff (appendix p 10). Likewise, as illustrated by the three-variable model using 0·5 as a cutoff, other metrics of clinical outcome, such as mortality at day 28 (nine [60%] *vs* six [14%]; p=0·0015) and ventilator-free days (0 days [IQR 0–2] *vs* 13 days [0–24]; p=0·0040), were also significantly worse in the hyperinflammatory phenotype. These findings in divergent outcomes were also observed with phenotypes derived using the four-variable model (data not shown). Significant differences in mortality were not observed when patients were stratified by their Acute Physiology, Age, and Chronic Health Evaluation (APACHE) III score (p=0·13).

Aside from clinical outcomes, plasma levels of several inflammatory biomarkers were higher in the hyperinflammatory phenotype compared with the hypoinflammatory phenotype and platelets were lower in the hyperinflammatory phenotype (figure 4). As with LCA-derived phenotypes, there was no significant difference in $PaO_2$–$FiO_2$ ratio between the identified phenotypes in START (figure 4).

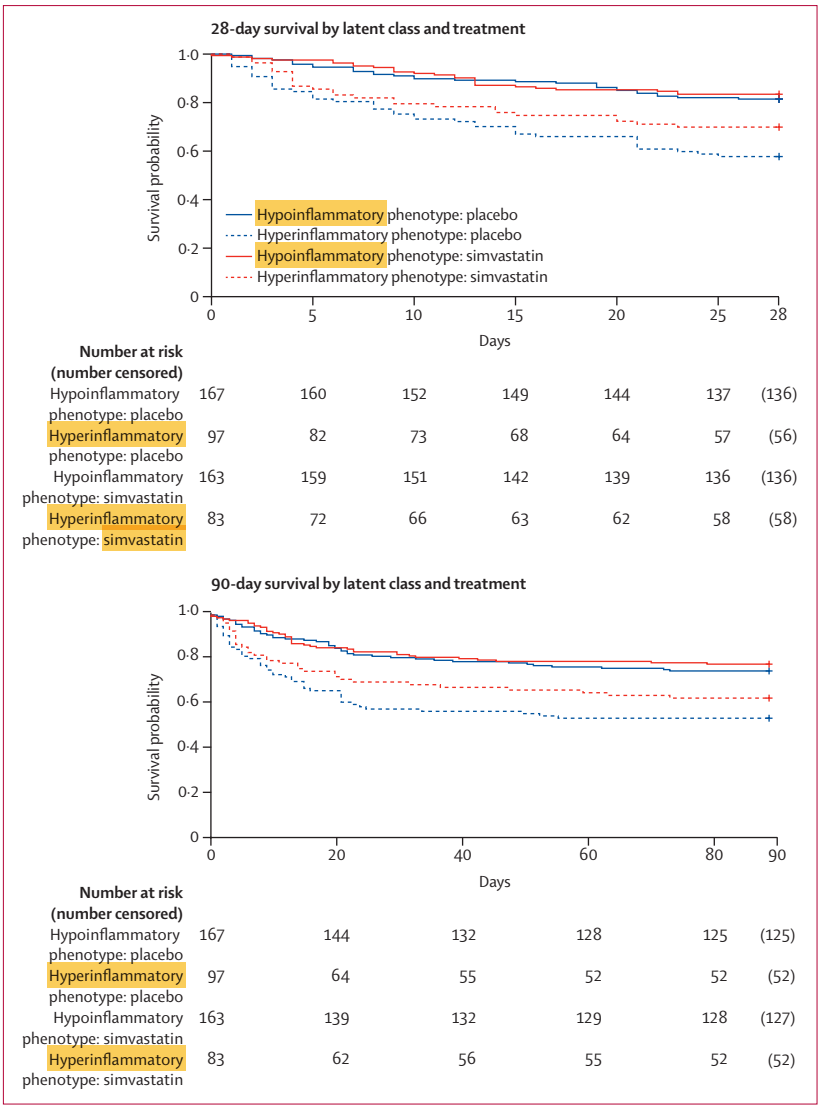



***Figure 3:*** **Kaplan-Meier survival curve in HARP-2 stratified by treatment and phenotypes assigned using a three-variable model**

The ancillary model included the variables interleukin-6, soluble TNF receptor 1, and vasopressor use. Class was assigned using a probability cutoff of ≥0·5 to assign phenotype. No patients were censored until the analysis endpoint, presented in brackets.

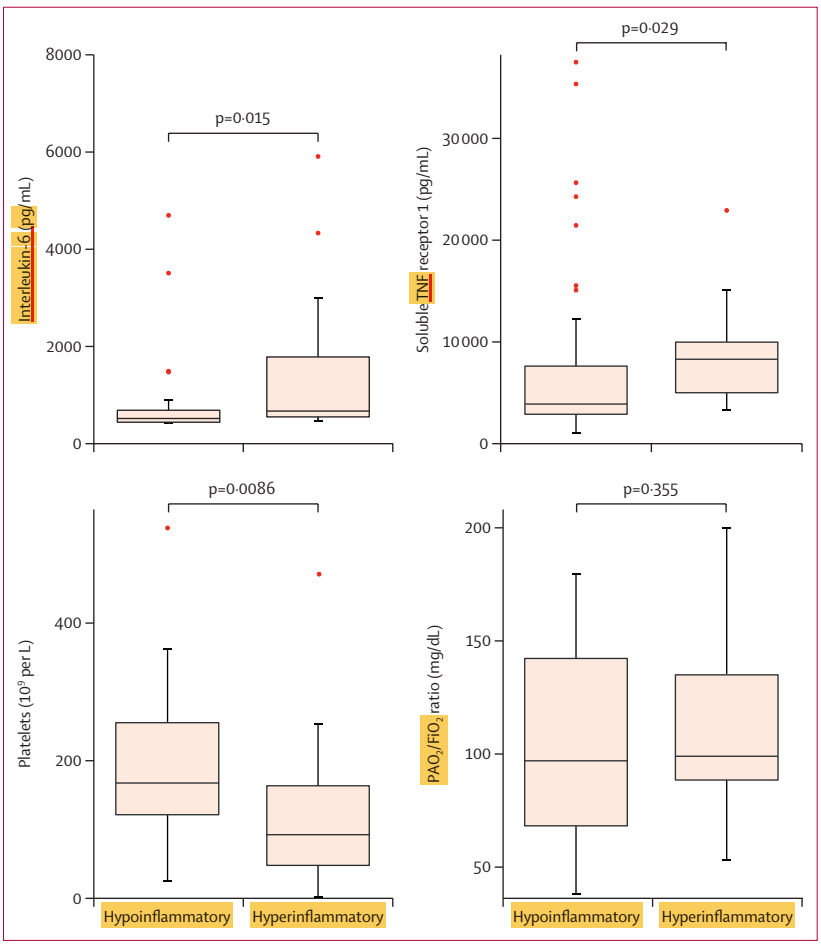



***Figure 4:*** **Difference in key variables between the hyperinflammatory and hypoinflammatory phenotypes in START external validation**

The three-variable model (interleukin-8, bicarbonate, and protein C) is used with a probability cutoff of ≥0·5 to assign phenotype. One value is not shown for interleukin-6 in the hypoinflammatory class due to *y*-axis censoring for visual interpretation. p values are representative of Mann-Whitney *U* test. $PaO_2/FiO_2$ ratio=ratio of partial pressure of oxygen in arterial blood to the fractional concentration of oxygen in inspired air. TNF=tumour necrosis factor.

## Discussion

LCA has consistently identified two ARDS phenotypes that show differential outcomes and response to treatment, but the complexity of LCA models has, to date, rendered ARDS phenotypes inaccessible in the clinical setting. In these analyses, parsimonious classifier models are presented that can accurately identify ARDS phenotypes. The ability to identify phenotypes using a small set of variables is a crucial step towards their clinical application and has important implications for the feasibility of future phenotype-guided clinical trials.

Elevated levels of pro-inflammatory cytokines, such as IL-8, IL-6, and soluble TNF receptor 1, are known individually to be associated with worse outcomes in ARDS and, unsurprisingly, emerged as the most important phenotype-defining variables. Protein C, a zymogen with anticoagulant and anti-inflammatory properties, was also an important variable, and lower levels have been independently associated with increased mortality and adverse outcomes in ARDS.[16]

Lower levels of bicarbonate in the setting of acute inflammation act as a surrogate for worsening metabolic acidosis, which in turn might reflect tissue hypoxia and dysregulated inflammation. Both protein C and bicarbonate, therefore, had negative coefficients in the models predicting the hyperinflammatory phenotype. Compared with previous studies that have used these variables in isolation to predict outcomes, the presented models developed and validated in this study have the additional benefit of using a composite of these variables and their values relative to each other.

Although the two best-fitting models both performed with high accuracy, the three-variable model (using IL-8, bicarbonate, and protein C) offers some obvious practical advantages for prospective clinical use. The added complexity of the four-variable model was insufficiently offset by additional accuracy. Moreover, the fourth variable, vasopressor use, is an ambiguous predictor variable. First, it does not factor in dose, thereby providing little insight into severity of shock. Second, the threshold to commence vasopressors in shock varies considerably and is often dictated by institutional, if not individual, discretion.[17] Therefore, the three-variable model that does not incorporate vasopressor use might be preferred. At the same time, the ability of the four-variable model itself, and vasopressor use independently, to identify patients with higher mortality[18] suggests that this model might be potentially valuable in certain ARDS trials.

A priori, a decision was made to compare two probability cutoffs to assign phenotype: the Youden Index from the derivation dataset and 0·5. In all models, the Youden Index cutoff was lower than the 0·5 cutoff and therefore unsurprisingly led to higher sensitivity but lower specificity. The proportion of patients with LCA-derived hyperinflammatory phenotype in the validation test set and HARP-2 was approximately 35%; this value was more closely matched when using 0·5 as the cutoff in all models. Calculating the Youden Index from the derivation (in-sample) dataset might have led to an overestimation of model accuracy.

In practical terms, the purpose of identifying phenotypes and the potential risk–benefit ratio of the proposed treatment strategy might ultimately dictate the best cutoff. For example, in a trial of a low-risk intervention, it could be reasonable to accept lower specificity to enhance sensitivity, whereas when studying a high-risk intervention, it might be more important to maximise specificity. Prospective studies are needed to further test optimal probability cutoffs.

In addition to the need for prospective validation, immediate implementation of these models is limited by the lack of a real-time test for biomarker quantification. To our knowledge, there are no commercially available point-of-care or real-time quantifiable assays for IL-8,

protein C, or soluble TNF receptor 1. The current study adds to the increasing weight of evidence that suggests that rapid measurement of plasma protein biomarkers could be crucial in delivering precision-based care in critical illness.[19] Recently, the National Heart, Lung, and Blood Institute convened a multidisciplinary working group to discuss the development of rapid biomarker testing in cardiovascular medicine.[20] Similar initiatives in critical care would be timely and essential to shift from the current over-reliance on a one-size-fits-all approach to treating syndromes such as sepsis and ARDS.

Keeping this limitation in mind, we adopted a pragmatic view on model development and sought to develop and evaluate ancillary models using permutations of the six most important variables. Most of these models were sufficiently accurate; however, those based on IL-6 were more sensitive and less specific for identifying the hyperinflammatory phenotype compared with IL-8-based models. One of the ancillary models afforded the opportunity to test its accuracy in the HARP-2 trial, where only a small set of variables were available for phenotyping. The ability of one of the least accurate ancillary models to not only identify phenotypes in HARP-2 but also to detect the disparate treatment effect in this dataset supports the robustness of the findings and their potential validity in trial cohorts beyond the ARDS Network.

The performance of the primary three-variable model in the START trial adds face validity to this argument. In START, albeit in a small cohort, the primary classifier models identified phenotypes that were distinct from each other and also had vastly divergent clinical outcomes. More importantly, when stratified by APACHE III score, the same differences in mortality were not observed in the phenotypes, suggesting that the severity of illness identified by phenotypes cannot be extracted from standard measures of severity. Pending rapid biomarker quantification, these models offer a simple and unique method for prognostic, and potentially predictive, enrichment.

This study has several strengths. First, we used four large RCT cohorts, where, to avoid overfitting, the validation cohort was kept completely naive to model development. Additionally, the derivation dataset was the largest in which we have applied LCA. The finding that the two-phenotype model was the optimal fit in this population suggests that ARDS phenotypes are consistent despite changing practice over two decades and across diverse populations. Second, the validation test set was a contemporary trial of infection-associated ARDS and had a higher incidence of the hyperinflammatory phenotype and significantly different levels of biomarkers and vasopressor use.

Despite these key differences in the derivation dataset and validation test set, the parsimonious regression models performed with high accuracy in the latter. Taken together with the model performance in the START and HARP-2 trials, the results suggest that the models are likely to be generalisable to other clinical trial populations in ARDS and robust to changes in assays and clinical practice over time, although prospective validation will still be required.

This study also has several limitations. All the presented data are secondary analyses of previously conducted RCTs. Interpretation of the performance of the parsimonious regression models must, therefore, be limited to trial populations. These models must be evaluated in observational cohorts and prospectively before they can be generalised to the ARDS population and used in the clinical setting. A further limitation is that all the studies used for this analysis except for the START trial were done before 2014. Since then, prone positioning has been shown to be beneficial in select populations of ARDS and is now in widespread use for some patients with severe ARDS.[2] We were unable to test the impact of prone positioning on phenotype allocation due to lack of data on this therapy. Additionally, the SAILS cohort represents a specific subset of patients with ARDS with infection or sepsis, albeit a subset that makes up the majority of patients with ARDS.

Another limitation was that the time from ARDS diagnosis to enrolment was different among cohorts (appendix p 5). This variability might have resulted in clinical management strategies playing an important, yet undetermined, role in patient phenotype. A previous study by our group has reported that phenotypes remained stable over a period of 72 h, suggesting little impact of management strategy on patient phenotype in this time frame.[21] Due to the retrospective study design, however, it is not feasible to ascertain the extent to which ventilatory and other management strategies leading up to enrolment altered the inflammatory response in these patients.

The limited set of variables available in HARP-2 meant that the accuracy of the two primary classifier models was not tested in this dataset. A further limitation of this study is that differential treatment response with phenotype assignment using a parsimonious regression model was only tested in HARP-2. Differential treatment responses in FACTT and ALVEOLI were not evaluated because both studies were included in the derivation dataset, and positive results would be subject to bias and data circularity.

What are some of the key knowledge gaps in the field going forward? Currently, identification of ARDS phenotypes using LCA has been limited to patients enrolled in RCTs, so it is unknown whether these phenotypes are generalisable to broader ARDS populations. Furthermore, it is also not known whether these phenotypes might be identifiable in critical care clinical syndromes beyond ARDS.

In particular, given that SAILS was a sepsis-associated cohort, these phenotypes might be applicable to sepsis. To fully realise the potential of these phenotypes to

deliver precision-based care in ARDS, a better understanding is needed of the underlying biology of the phenotypes; this objective will require more experimental research. Additionally, a better understanding of the longitudinal kinetics of the phenotypes and their response to interventions is needed. For example, the diagnosis of ARDS itself is known to be volatile to standard ventilatory practice in the first 24 h;[22] whether these changes are specific to either phenotype or affect phenotype assignment themselves is unknown.

In summary, this study provides evidence for accurate parsimonious classifier models for ARDS phenotypes. These simple models might facilitate the study of phenotypes in the prospective setting and improve selection of patients for clinical trials.

#### Contributors

PS, KLD, CMO'K, DFM, MAM, and CSC all conceived and designed the study. CMO'K, DFM, and MAM collected the data. PS, KLD, MAM, and CSC did the data analysis and interpretation. PS, KLD, and CSC drafted the manuscript. All authors contributed to revisions of the manuscript. The final version of the manuscript was read and approved by all authors.

#### Declaration of interests

PS was supported by the National Institutes of Health (NIH) during the period that the study was conducted. CSC reports grants from the NIH, during the conduct of the study; and grants from GlaxoSmithKline; grants and personal fees from Bayer; and personal fees from Prometic, Roche/Genentech, CSL Behring, and Quark, outside of the submitted work. CMO'K reports grants from the National Institute for Health Research (NIHR) Efficacy and Mechanism Evaluation (EME) during the conduct of the study. In addition, she has received grant funding from the NIHR, Wellcome Trust, Northern Ireland Health and Social Care Research and Development, and other funders for ARDS studies. CMO'K reports her spouse has received consultancy fees from GlaxoSmithKline, Bayer, and Boehringer Ingelheim outside of the submitted work. DFM reports grants from NIHR EME, Health Research Board, Northern Ireland Public Health Agency Research and Development, Intensive Care Society of Ireland, and REVIVE for the conduct of the HARP-2 study. DFM reports personal fees from consultancy for GlaxoSmithKline, Boehringer Ingelheim, and Bayer, outside of the submitted work. In addition, his institution has received funds from grants from the UK NIHR, Wellcome Trust, Innovate UK, and others. In addition, DFM has a patent issued to his institution for a treatment for ARDS. DFM is a Director of Research for the Intensive Care Society and NIHR EME Programme Director. MAM reports grants from NIH/National Heart, Lung, and Blood Institute, the US Department of Defense, Bayer Pharmaceuticals, and GlaxoSmithKline; and personal fees from Cerus Therapeutics, outside of the submitted work. KLD declares no competing interests.

#### Acknowledgments

This study was funded by the National Institutes of Health, grant numbers HL131621, HL133390, HL140026 (CSC), and GM008440-21 (PS). We thank all patients who participated in the National Heart Lung and Blood Institute (NHLBI) ARDS Network trials from which data from this study were derived. These include the ARMA, ALVEOLI, FACTT, and SAILS trials. We acknowledge the contributions of health-care providers and research staff that enabled the successful completion of these trials. In addition, we thank the contributions of the Biological Specimen and Data Respository Information Coordinating Center of the NHLBI that made the data and biological specimens available to do these studies. We thank all patients who participated in the HARP-2 trial and their legal representatives, all research nurses and pharmacists in the participating centres, and medical and nursing staff in participating centres who cared for patients and collected data. For the HARP-2 trial, we thank the staff of the Northern Ireland Clinical Trials Unit for their support in conducting the trial, the staff from the Health Research Board Galway Clinical Research Facility (Galway, Ireland) for help in conducting the study in Ireland, the staff of the Northern Ireland Clinical Research Network and the NIHR Clinical Research Network for help with patient recruitment and data acquisition, and the UK Intensive Care Society. Work on the HARP-2 study was supported by the UK Efficacy and Mechanism Evaluation Programme, a Medical Research Council (MRC) and National Institute for Health Research (NIHR) partnership (08/99/08 and 16/33/01), by a Health Research Award (HRA_POR-2010–131) from the Health Research Board, Dublin; and by additional funding from the Health and Social Care (HSC) Research and Development Division, Public Health Agency for Northern Ireland, the Intensive Care Society of Ireland, and REVIVE. The EME Programme is funded by the MRC and NIHR, with contributions from the Chief Scientist Office in Scotland, the National Institute for Social Care and Health Research in Wales, and the HSC Research and Development Division, Public Health Agency for Northern Ireland. The views expressed in this article are those of the authors and not necessarily those of the MRC, National Health Service, NIHR, or Department of Health.

#### References

1 Acute Respiratory Distress Syndrome Network, Brower RG, Matthay MA, et al. Ventilation with lower tidal volumes as compared with traditional tidal volumes for acute lung injury and the acute respiratory distress syndrome. *N Engl J Med* 2000; **342:** 1301–08.

2 Guerin C, Reignier J, Richard JC, et al. Prone positioning in severe acute respiratory distress syndrome. *N Engl J Med* 2013; **368:** 2159–68.

3 National Heart, Lung, and Blood Institute Acute Respiratory Distress Syndrome Clinical Trials Network, Wiedemann HP, Wheelper AP, et al. Comparison of two fluid-management strategies in acute lung injury. *N Engl J Med* 2006; **354:** 2564–75.

4 Matthay MA, McAuley DF, Ware LB. Clinical trials in acute respiratory distress syndrome: challenges and opportunities. *Lancet Respir Med* 2017; **5:** 524–34.

5 Calfee CS, Delucchi K, Parsons PE, et al. Subphenotypes in acute respiratory distress syndrome: latent class analysis of data from two randomised controlled trials. *Lancet Respir Med* 2014; **2:** 611–20.

6 Famous KR, Delucchi K, Ware LB, et al. Acute respiratory distress syndrome subphenotypes respond differently to randomized fluid management strategy. *Am J Respir Crit Care Med* 2017; **195:** 331–38.

7 Calfee CS, Delucchi KL, Sinha P, et al. Acute respiratory distress syndrome subphenotypes and differential response to simvastatin: secondary analysis of a randomised controlled trial. *Lancet Respir Med* 2018; **6:** 691–98.

8 Sinha P, Delucchi KL, Thompson BT, et al. Latent class analysis of ARDS subphenotypes: a secondary analysis of the statins for acutely injured lungs from sepsis (SAILS) study. *Intensive Care Med* 2018; **44:** 1859–69.

9 Villar J, Ambros A, Mosteiro F, et al. A prognostic enrichment strategy for selection of patients with acute respiratory distress syndrome in clinical trials. *Crit Care Med* 2019; **47:** 377–85.

10 Sinha P, Calfee CS, Beitler JR, et al. Physiologic analysis and clinical performance of the ventilatory ratio in acute respiratory distress syndrome. *Am J Respir Crit Care Med* 2019; **199:** 333–41.

11 McAuley DF, Laffey JG, O'Kane CM, et al. Simvastatin in the acute respiratory distress syndrome. *N Engl J Med* 2014; **371:** 1695–703.

12 Brower RG, Lanken PN, MacIntyre N, et al. Higher versus lower positive end-expiratory pressures in patients with the acute respiratory distress syndrome. *N Engl J Med* 2004; **351:** 327–36.

13 National Heart, Lung, and Blood Institute Acute Respiratory Distress Syndrome Clinical Trials Network, Truwit JD, Bernard GR, et al. Rosuvastatin for sepsis-associated acute respiratory distress syndrome. *N Engl J Med* 2014; **370:** 2191–200.

14 Matthay MA, Calfee CS, Zhuo H, et al. Treatment with allogeneic mesenchymal stromal cells for moderate to severe acute respiratory distress syndrome (START study): a randomised phase 2a safety trial. *Lancet Respir Med* 2019; **7:** 154–62.

15 Altmann A, Tolosi L, Sander O, Lengauer T. Permutation importance: a corrected feature importance measure. *Bioinformatics* 2010; **26:** 1340–47.

16 Terpstra ML, Aman J, van Nieuw Amerongen GP, Groeneveld AB. Plasma biomarkers for acute respiratory distress syndrome: a systematic review and meta-analysis. *Crit Care Med* 2014; **42:** 691–700.

17 Bangash MN, Kong ML, Pearse RM. Use of inotropes and vasopressor agents in critically ill patients. *Br J Pharmacol* 2012; **165:** 2015–33.

18 National Heart, Lung, and Blood Institute Acute Respiratory Distress Syndrome Clinical Trials Network, Matthay MA, Brower RG, et al. Randomized, placebo-controlled clinical trial of an aerosolized $\beta_2$-agonist for treatment of acute lung injury. *Am J Respir Crit Care Med* 2011; **184:** 561–68.

19 Garcia-Laorden MI, Lorente JA, Flores C, Slutsky AS, Villar J. Biomarkers for the acute respiratory distress syndrome: how to make the diagnosis more precise. *Ann Transl Med* 2017; **5:** 283.

20 King K, Grazette LP, Paltoo DN, et al. Point-of-care technologies for precision cardiovascular care and clinical research: National Heart, Lung, and Blood Institute Working Group. *JACC Basic Transl Sci* 2016; **1:** 73–86.

21 Delucchi K, Famous KR, Ware LB, et al. Stability of ARDS subphenotypes over time in two randomised controlled trials. *Thorax* 2018; **73:** 439–45.

22 Villar J, Perez-Mendez L, Lopez J, et al. An early PEEP/$FiO_2$ trial identifies different degrees of lung injury in patients with acute respiratory distress syndrome. *Am J Respir Crit Care Med* 2007; **176:** 795–804.

THE LANCET
Respiratory Medicine

# Supplementary appendix

This appendix formed part of the original submission and has been peer reviewed. We post it as supplied by the authors.

Supplement to: Sinha P, Delucchi KL, McAuley DF, O'Kane CM, Matthay MA, Calfee CS. Development and validation of parsimonious algorithms to classify acute respiratory distress syndrome phenotypes: a secondary analysis of randomised controlled trials. *Lancet Respir Med* 2020; published online Jan 13. http://dx.doi.org/10.1016/S2213-2600(19)30369-8.

## 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 training 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.