ERJ open research
THE BEST IN OPEN ACCESS BASIC, TRANSLATIONAL & CLINICAL RESPIRATORY RESEARCH

Early View

Original research article

Computational platform for doctor-AI cooperation in PAH prognostication: a pilot study

Vitaly O. Kheyfets, Andrew J. Sweatt, Mardi Gomberg-Maitland, Dunbar D. Ivy, Robin Condliffe, David G. Kiely, Allan Lawrie, Bradley A. Maron, Roham T. Zamanian, Kurt R. Stenmark

Please cite this article as: Kheyfets VO, Sweatt AJ, Gomberg-Maitland M, *et al*. Computational platform for doctor-AI cooperation in PAH prognostication: a pilot study. *ERJ Open Res* 2022; in press (https://doi.org/10.1183/23120541.00484-2022).

Computational Platform for Doctor-AI Cooperation in PAH Prognostication: A pilot study

Running Title: A platform for doctor-algorithm cooperation in PAH

Vitaly O. Kheyfets PhD[1*], Andrew J. Sweatt MD[2,3], Mardi Gomberg-Maitland MD[4], Dunbar D. Ivy MD[5], Robin Condliffe MD[6], David G. Kiely MD[6,7,8], Allan Lawrie PhD[6,7,8], Bradley A. Maron MD[9], Roham T. Zamanian MD[2,3], Kurt R. Stenmark MD[1]

[1] Paediatric Critical Care Medicine; Developmental Lung Biology and CVP Research Laboratories, School of Medicine, University of Colorado (Aurora, CO, USA).

[2] Division of Pulmonary and Critical Care Medicine, Stanford University (Stanford, CA, USA)

[3] Vera Moulton Wall Center for Pulmonary Vascular Disease, Stanford University (Stanford, CA, USA)

[4] Division of Cardiology, George Washington University Hospital (Washington, DC, USA)

[5] Department of Paediatric Cardiology, Children's Hospital Colorado (Aurora, CO, USA)

[6] Sheffield Pulmonary Vascular Disease Unit, Sheffield Teaching Hospitals NHS Foundation Trust, Royal Hallamshire Hospital (Sheffield, UK)

[7] Department of Infection, Immunity and Cardiovascular Disease, University of Sheffield (Sheffield, UK)

[8] Insigneo Institute for in-silico Medicine, University of Sheffield (Sheffield, UK)

[9] Division of Cardiovascular Medicine, Brigham and Women's Hospital and Harvard Medical School, Harvard University (Boston, MA, USA).

Preprint Request and Correspondence
Vitaly O. Kheyfets
University of Colorado Anschutz Medical Campus, School of Medicine
Research Complex 2 - 12705 E. Montview BLVD, Office 6111
Aurora, CO 80045
Email: vitaly.kheyfets@cuanschutz.edu
Office Phone: 303-724-9764 / Mobile Phone: 913-568-4408

Take Home Message:
High throughput biomarker screening and machine learning (ML) are promising new technologies that could revolutionize the way doctors screen PAH patients. We show how principles of game theory combined with ML modeling would allow doctor-ML collaboration.

Word count: 2990

ABSTRACT

Background: Pulmonary arterial hypertension (PAH) is a heterogenous and complex pulmonary vascular disease associated with substantial morbidity. Machine learning algorithms (used in many PAH risk calculators) can combine established parameters with thousands of circulating biomarkers to optimize PAH prognostication, but these approaches do not offer the clinician insight into what parameters drove the prognosis. The approach proposed in this study diverges from other contemporary phenotyping methods by identifying patient-specific parameters driving clinical risk.

Methods: We trained a random forest (RF) algorithm to predict 4-year survival risk in a cohort of 167 adult PAH patients evaluated at Stanford university, with 20% withheld for (internal) validation. Another cohort of 38 patients from Sheffield university were used as a secondary (external) validation. Shapley values, borrowed from game theory, were computed to rank the input parameters based on their importance to the predicted risk score for the entire trained RF model (global importance) and for an individual patient (local importance).

Results: Between the internal and external validation cohorts, the RF model predicted 4-year risk of death/transplant with a sensitivity and specificity between 71.0-100% and 81.0-89.0%, respectively. The model reinforced the importance of established prognostic markers, but also identified novel inflammatory biomarkers that predict risk in some PAH patients.

Conclusion: These results stress the need for advancing individualized phenotyping strategies that integrate clinical and biochemical data with outcome. The computational platform presented in this study offers a critical step towards personalized medicine in which a clinician can interpret an algorithm's assessment of an individual patient.

Number of words: 250

Key Words: Pulmonary hypertension, outcomes calculator, machine learning

INTRODUCTION

Pulmonary arterial hypertension (PAH) is a highly morbid disease characterized by complex pathobiology and variable clinical presentation [1]. The availability of numerous approved [2] and emergent [3] PAH medications has expanded treatment options used clinically to limit morbidity and prevent premature death. However, pharmacotherapeutic initiation and dose escalation is guided by risk assessment [4], emphasizing the importance of data and algorithms that clarify prognosis in individual patients. In turn, pathogenetic and phenotypic heterogeneity across the PAH spectrum introduce unique challenges to precision-based risk stratification methods.

Several validated risk estimation tools are now available for PAH and have transformed clinical decision-making by allowing evidence-based prognostication at point-of-care [5, 6], but these algorithms consider a relatively narrow range of variables that can't account for disease heterogeneity [7-9]. Utilizing machine learning algorithms allows contemporary calculators to compute risk estimates from robust datasets that can include thousands of circulating biomarkers [10, 11] across diverse PAH populations, thereby expanding the gamut of patient-specific measurements available for compiling risk estimates. However, a clinician utilizing these complex multi-variate models on a patient would only be presented with a numeric risk score without any explanation of how the model reached its calculation. Therefore, these models require an analytic strategy to generate an intuitive readout that explains how each marker contributed to the final prediction, which would mark a significant advance towards individualizing clinical decision-making in PAH.

The overall goal of this project was to showcase a computational platform that i) integrates big data inclusive of biological and clinical parameters to generate a composite risk profile and

ii) ranks the contribution of all patient parameters to the computed risk score for an individual patient (using game theory). Combining these two mathematical principals for risk stratification is positioned to advance the use of artificial intelligence for personalized clinical-decision making in PAH.

METHODS

To demonstrate the aforementioned computational platform for doctor-machine interaction in PAH prognostication, we utilized the dataset presented in ref. [7] (referred to here as the *Case Study dataset*) to train an RF model that predicts the probability of death/transplant within 4-years. Once a suitably accurate model is developed, we then applied game theory principles [12] to explain overall model structure and showcase how a clinician could interpret the prediction for a specific patient.

Case Study Dataset - Study population and design:

The study analyzed data from a prospective observational cohort of 281 PAH patients evaluated at Stanford University (Stanford, CA) who had banked peripheral blood samples between 2008 and 2014. This dataset has been extensively described in ref. [7]. Within this cohort, 114 patients were known to be alive and transplant-free before the 4-year follow-up, but not evaluated beyond this point. Therefore, this study considers 167 patients (demographics shown in Table 2) with a documented 4-year outcome (N = 93 transplant-free survival vs. N = 74 death or lung transplantation), where the 4-year cut-off was chosen to achieve a roughly 50% event rate because machine learning algorithms tend to produce erroneous classifiers when trained on imbalanced datasets.

For each patient, the model was trained on 93 patient parameters (Fig. 1a; all variables defined in Table 1) that included: (1) demographics; (2) PAH subgroup; (3) functional metrics and standard clinical bloodwork; (4) invasive hemodynamic measurements; (5) lung function measurements; (6) selected echocardiographic parameters; and (7) an exploratory proteomic immune panel of 48 cytokines, chemokines, and growth factors measured using a Bio-Plex multiplex immunoassay (Bio-Rad Inc, Hercules, CA) (see Fig. 1b).

Feature Engineering

Of the 93 patient parameters, 4 categorical parameters (sex; ethnicity; PAH subgroup; and presence of pericardial effusion) were one-hot-encoded (i.e., coding categorical variables as binary vectors) [13] to arrive at a dataset with 167 observations and 104 features.

The dataset had 863 (<5% of total data matrix) missing values. For continuous variables that were distributed normally, each missing value was replaced with the feature mean. For categorical variables or continuous variables that were not distributed normally, the missing values were replaced with the feature mode.

To reduce the dimension of the original dataset, we performed recursive feature elimination (RFE) (*viz.*, 10,000 trees trained at each iteration; each tree is allowed 4 decision splits; 3% of the least important features are removed at each iteration) as is described in ref. [14] (using Matlab 2020b, Mathworks). The variable set producing the lowest out-of-bag classification error [15] contained 23 features, which was too high for a predictive model with 167 observations. Therefore, we allowed the RFE algorithm to run until less than 10 features remained, thus producing a final data matrix with 9 features.

Developing the final predictive model and validation:

A final RF model (with 10,000 trees based on plateaued out-of-bag classification error; each tree was allowed to grow to a maximum depth of 4 generations) was trained on 80% of the observations, with 20% withheld for (internal) validation (using "sklearn" library [13] in Python 3.9.6). An additional external validation cohort of 38 PAH patients from the University of Sheffield, collected from the Sheffield Pulmonary Vascular Disease Unit between 2008 and 2014, was also utilized to assess model performance. Therefore, the results of this study will reference the *internal validation cohort* (20% of the Stanford cohort -see Table 2- withheld for testing) and the *external validation cohort* (the Sheffield cohort, see Online Supplementary Material O2, trained on 100% of the Stanford patients).

Because the overall objective of this study was to showcase model interpretability, and there is no point in interpreting a model that is not sufficiently accurate, we compare the performance of the final trained RF model against the REVEAL 2.0 calculator [6] as a standard for sufficient model accuracy. All sensitivity/specificity values reported in the manuscript were computed by generating a receiver operating curve (ROC) for computed probability values (when evaluating the RF model) or REVEAL 2.0 risk score.

RF model interpretability:

In 1951, Lloyd Shapley built on the concept of Game Theory by deriving an equation for the marginal contribution of a single player in a cooperative game [16], which won him the Nobel Prize in economics. The Shapley value is computed for each player in a cooperative game to fairly determine the marginal contribution of that player, which -through interaction between players- might be different than the player's individual score. This concept has been expanded to

"explain" the marginal contribution of each feature in multivariate tree-based machine learning (ML) models [12], which allows for both global (how input features rank and contribute to the overall model prediction) and local (how input features rank and contribute to an individual model prediction) interpretability of feature contribution and interaction.

SHAP values and the reported plots that use them were computed with the *TreeExplainer* SHAP package in Python [12]. Even though the direct computation of Shapley values would be too computationally expensive, especially as the number of considered features would be increased, the computational pipeline outlined in ref. [12] is fast for even high-dimensional problems, and guarantees "local accuracy" and "consistency" [17]. Global model structure was explored using violin plots. Local model interpretability (defined as the degree that a human can understand the cause of a model's decision), used to understand the model prediction for a specific patient, was assessed using violin plots and decision plots.

RESULTS

Random Forest Model of Risk (a case study):

After performing RFE, the final model consisted of 9 features: (1) 6MWD; (2) DLCO; (3) NT-proBNP; (4) Lymph (%); (5) Lymph (abs); (6) IL-9; (7) IL-2; (8) SCF; and (9) HGF, which all had significantly different means between high risk and low risk patients (see. Fig. 2, see Table 1 for acronym reference). The area under the ROC (AUC) shown in Fig. 2 revealed that all 9 markers fairly discriminated high vs. low-risk patients, but the discrimination accuracy is considerably improved by combining all 9 into a multivariate model (see Fig. 3). It's critical to note that we are not suggesting for these 9 features to be clinically utilized based on this relatively small patient cohort. Although we are encouraged by the fact that our RFE algorithm

identified markers commonly accepted as prognostic, these metrics would need to be validated in larger prospective studies. An expanded discussion of these results is available in the Online Supplement O1.

For the internal validation cohort (see Fig. 3b), the RF model accurately predicted 15 of 17 patients as high risk. The model's computed probability of 4-year all-cause mortality risk in the internal validation cohort produced an AUC of 0.94 (95% confidence interval, CI = 0.79-1.00), with a sensitivity and specificity of the RF model of 1.00 and 0.89, respectively (see Fig. 3c). Pointwise confidence intervals on the sensitivity were computed using vertical averaging from 1000 sampled bootstrap replicas. For the external validation cohort, which used an estimate of the 6MWD and serum NT-proBNP measurements (see Online Supplement O2), the model revealed an AUC of 0.81 (95% CI = 0.64-0.92) and sensitivity and specificity of 0.71 and 0.81, respectively (see Fig. 3c).

Global and local Interpretability of the Random Forest Model:

Given that the RF model has 10,000 trained decision trees, it's not possible to intuitively understand what parameters are driving model prediction and how they influence the overall computed score, which is referred to as the "global model structure." While there have been numerous methods proposed for evaluating global model structure (e.g., ranking feature importance [18]), our approach can also be applied to an individual patient (described in the next section).

Fig 4 shows a violin summary plot of SHAP values for the training (Fig. 4a) and internal validation (Fig. 4b) datasets. The top 3 features are the same in both datasets, thus suggesting a consistent global model structure between the training and validation cohorts. The features listed

are those identified as “most important” through RFE and ordered along the vertical axis based on global feature importance in the RF model. A single dot represents a patient, and the width of the violin is representative of the number of patients that fall into that region. As an example, based on the limited case study presented here, the violin plots shown in Fig. 4 allows us to conclude that a high 6MWD can reduce the probability of death in 4-years by up to 20%, but the long tail towards positive SHAP values seen for IL2 and IL9 would suggest that, even though they are at the bottom of the global importance plot, these cytokine levels could be extremely important for certain individuals.

Local Interpretability of the trained RF Model for a specific patient:

When asking a trained RF model to make a risk prediction for a new incoming patient, the algorithm runs that new patient’s data through the 10,000 trees that were generated during the training process. Each tree classifies the patient as high or low-risk, and the algorithm then takes a majority vote to make its prognosis. The numerical divide of how each tree voted also offers an estimate of probability, although a calibration plot showed that, for the case study dataset considered here, the model probability values were overly conservative and unresponsive to Platt’s scaling [19], which was likely due to a small validation cohort. Therefore, if a clinician is interested in knowing how each patient measurement contributed to the risk score, it is again not possible for a human to make sense of this prediction from 10,000 decision trees.

Fig 5a shows decision plots for the internal validation cohort with dashed lines representing the two patients who were incorrectly predicted to have a high 4-year risk of mortality (note that the feature order is identical to Fig. 4b). The vertical axis lists the features in

order of importance for that specific cohort of patients, so the order might be different if some patients were removed or if only one patient was being considered.

Decision plots are used to show how a model -for any individual patient or for combined patients- reaches the predicted 4-year risk score. All decision tree lines start at the same point (at a risk score of 0.55 in Fig. 5a) along the bottom horizontal axis, which represents the baseline predicted score before any of the features were considered. The SHAP values (i.e., the change in the score in response to a specific feature) accumulate from the base value to arrive at the RF model's final score on the top horizontal axis.

Case example 1. In Fig. 5a, we focus on two random patients, indicated by arrows. Both patients had a relatively normal 6MWD and NT-proBNP and, therefore, based on the current approach to risk stratification, might be expected to harbour similar risk profiles. However, patient 1 had abnormally low IL2 and IL9 levels based on comparisons in ref. [7] (also see Fig. 2), which in spite of normal exercise tolerance and markers of heart failure, put them firmly in the high-risk group. Alternatively, all 9 metrics for patient 2 were in the normal range, thus resulting in assignment to the low risk group.

Fig. 5b shows a decision tree for the internal validation cohort that considers the cumulative effect of first-order interactions between the features. Here we see that interactions can certainly drive the final risk score for some patients but are all ranked as the "least important" features for the model output when the entire internal validation cohort is considered. For this reason, and because considering interactions significantly complicate the interpretability of the final model, we omit them for single-patient analysis in Fig. 6.

Case example 2. Fig. 6 (a and b) shows decision plots for two randomly selected patients from the internal validation cohort. Here we see that both patients were assigned a NYHA-FC=1 with a similar REVEAL 2.0 risk score, but Patient 1 had a mortality event within 4 years of study enrolment. The RF model trained in this case study accurately predicted Patient 1 to be in the high-risk group, driven primarily by circulating levels of IL2 and IL9, despite low-risk 6MWD and NT-proBNP. The profile for patient 2 included all 9 markers within the low-risk range, corresponding to correct assigned to the low-risk group.

DISCUSSION

Prior reports using trained machine learning models (e.g., RF models) in PAH have been effective for identifying biomarkers that contribute to risk estimation across patient cohorts [20-22]. However, these algorithms produce a single classification or score when asked to make a prediction for a specific patient, so the clinician might be hesitant to make treatment decisions based on "black box" predictions. Therefore, if the clinician can interact with the algorithm and is graphically presented (using decision plots) how the parameters are ranked in the final decision, that clinician would be better suited to generate a treatment strategy based on the algorithm's assessment or possibly overrule the algorithm based on their own clinical intuition.

In this paper, we use a case-study to introduce a platform that generates a graphical explanation of the RF model's prognosis for an individual PAH patient (a step towards personalized medicine). Because the cohort available for our case study had a limited number of patients for model training, we utilized RFE to reduce the 93 available patient measurements to 9. We then showed in two validation cohorts that our RF model's accuracy was comparable to the REVEAL 2.0 calculator (an expanded discussion on comparing against the REVEAL 2.0 calculator is available in Online Supplement O3).

Global interpretability studies showed that the general structure of the trained RF model heavily skewed towards known prognostic markers of PAH (e.g., exercise tolerance and heart failure). However, the violin plots for most of the circulating inflammatory markers in Fig. 4 had long tails, thus suggesting that they could also be a major driver of the prognostic score for some patients. This is confirmed by decision plots for the entire internal validation cohort (see Fig. 5), which showed that -consistent with multiple previous studies [10, 11, 23, 24]- inflammatory markers and their occasional interaction with other metrics can heavily influence the prognosis. This is also seen when looking at individual decision plots from randomly selected patients in Fig. 6. As an example, Patient 1 was placed in a low-risk category based on NYHA-FC and REVEAL 2.0 score, which don't consider biomarkers of inflammation. However, the RF model -trained with circulating markers of inflammation- correctly classified this patient as high-risk and the decision plot shows that it was the decreased levels of circulating cytokines that drove that prognosis. Interestingly, both IL2 and many other inflammatory cytokines are known to be upregulated in PAH patients, relative to controls (CTLs) [7, 25-27], but are reduced in PAH patients with poor survival. This would suggest that the presence of these markers could be protective and that the time-course of cytokine levels is itself prognostic, but this would need to be explored in future studies.

A major limitation of the current study was the relatively modest cohort available for model training and (internal and external) validation. This also prevented us from evaluating if our model was sufficiently calibrated because calibration curves require a large number of samples [28]. Future studies will re-evaluate our results in a larger cohort, but here we focused on showcasing the computational pipeline for doctor-algorithm interaction.

The external validation utilized in our study offers both a strength and a weakness. We were encouraged to find that -even though 6MWD was estimated from MSWT and NT-proBNP was measured in serum– the model still performed reasonably well. However, this inconsistency required us to focus on an internal validation dataset within the main manuscript, which can suffer from the same confounding biases as the training cohort and present an overinflated view of model performance.

Conclusion:

In this study, we present a novel computational pipeline for clinician-algorithm interaction in PAH risk assessment using a case study of prospectively analyzed patients. This approach can be expanded to consider hundreds and even thousands of patient measurements, thus introducing a critical step towards implementing big data and artificial intelligence into clinical decision making and entering the era of personalized medicine.

Funding - V.O.K. was supported by a NIH/NHLBI K25 career development award (5K25 HL 133481). A.J.S. was supported by a NIH/NHLBI K23 career development award (5K23HL15189202). This work was also supported by NIH P01HL152961 and P01HL01498.

Conflict of Interest – None.

TABLES

Variable	*Description*	*Units*
Age		Years
6MWD	Six-minute walk distance	Meters
MSWT	Modified shuttle walk test distance	Meters
e6MWD	$= 0.7225 \cdot MSWT + 70.769$, estimated 6MWD	Meters
NT-proBNP	Circulating N-terminal B-type natriuretic peptide	pg/mL
Creatine	Serum creatine concentration	pg/mL
Glomerular	Glomerular filtration rate	mL/min/1.73 m^2
WBC	CBC white blood cell count	10^3 cells/mm^3
Lymph (abs)	CBC differential absolute lymphocytes count	10^3 cells/mm^3
Lymph (%)	CBC differential percent lymphocytes count amongst all WBCs	%
Mono (abs)	CBC differential absolute monocyte count	10^3 cells/mm^3
Mono (%)	CBC differential percent monocyte count amongst all WBCs	%
Neut (abs)	CBC differential absolute neutrophil count	10^3 cells/mm^3
Neut (%)	CBC differential percent neutrophil count amongst all WBCs	%
Baso (abs)	CBC differential absolute basophil count	10^3 cells/mm^3
Baso (%)	CBC differential percent basophil count amongst all WBCs	%
Eos (abs)	CBC differential absolute eosinophil count	10^3 cells/mm^3
Eos (%)	CBC differential percent eosinophil count amongst all WBCs	%
FVC	Forced vital capacity during PFT	%
FEV1	Forced expiratory volume in 1-second during PFT	%
TLC	Total lung capacity during PFT within	%
DLCO	Diffusion capacity of lung for carbon monoxide during PFT	%
DLCO hemo	DLCO adjusted for hemoglobin during PFT	%
Effusion	Pericardial effusion observed at TTE	mm
RVFAC	RV fractional area change at TTE	%
TAPSE	Tricuspid annular plane systolic excursion	cm
RVPsys_fTR	RV systolic pressure computed from tricuspid regurgitation	mmHg
RVOT_VTI	RV outflow tract velocity time integral at TTE	cm/s
sPAP	Systolic pulmonary arterial pressure	mmHg
dPAP	Diastolic pulmonary arterial pressure	mmHg
mPAP	Mean pulmonary arterial pressure	mmHg
RAP	Right atrial pressure	mmHg
PCWP	Pulmonary capillary wedge pressure	mmHg
RVEDP	RV end diastolic pressure	mmHg
HR	Heart Rate	beats/min
CO	Cardiac output	mL/s
PVR	Pulmonary Vascular Resistance	Wood units
BSA	Body surface area	m^2
SVR-index	Systemic vascular resistance index	(Wood units2)* m^2
Compliance	= (sPAP-dPAP)/(stroke volume)	mmHg/mL
HGF	Hepatocyte growth factor	AU
SCF	Stem cell factor	AU
IL2	Interleukin 2	AU
IL9	Interleukin 9	AU

Table 1. Table of variables used in this study. Notation: CBC = complete blood count; PFT = pulmonary function test within 3-months; TTE = transthoracic echocardiogram; RV = right ventricle.

	H4Y-risk	L4Y-risk	$P_{2\text{-tailed}}$
Sample Size	74	93	
Age (years ± SD)	52.7 ± 15.5	47.0 ± 14.3	0.015
% Female Patients	71.6	75.3	
Race			
White	42	53	0.992
Asian	10	24	
Hispanic	11	11	
Black	5	3	
Other	6	2	
PAH subtype			
Connective Tissue Disease	25	27	0.997
Idiopathic PAH	19	27	
Drug and Toxins	13	15	
Congenital Heart Disease	8	17	
Portopulmonary Hypertension	8	4	
Hereditary PAH	1	3	
NYHA Functional Class			
Class I	4	5	0.942
Class II	11	42	
Class III	46	37	
Class IV	13	9	
Hemodynamics			
mPAP (mmHg)	50.8 ± 16.2	50.9 ± 16.6	0.969
PVR (dyn·s/cm^5)	11.2 ± 7.14	11.3 ± 6.59	0.925
Cardiac Index (m^2·min)	2.27 ± 0.82	2.24 ± 0.67	0.795
Mean Right Atrial Pressure (mmHg)	9.70 ± 6.33	7.89 ± 4.78	0.037
PCWP (mmHg)	12.2 ± 5.72	10.5 ± 4.09	0.027
***Timing* from …**			
Diagnosis (years ± SD)	3.1 ± 3.9	4.5 ± 5.4	0.053
Symptom Onset (years ± SD)	4.4 ± 4.8	6.0 ± 5.4	0.051
Therapy			
Treatment Naive	23	28	0.999
Monotherapy	21	28	
Dual Therapy	22	26	
Triple Therapy	8	11	

Table 2. Stanford cohort patient characteristics in high 4-year risk (H4Y-risk) and low 4-year risk (L4Y-risk) groups. Continuous data is compared using a t-test and categorical variables by χ^2 test.

REFERENCES

[1] Mandras, S. A., Mehta, H. S., and Vaidya, A., 2020, "Pulmonary Hypertension: A Brief Guide for Clinicians," Mayo Clinic Proceedings, 95(9), pp. 1978-1988.

[2] Maron, B. A., Abman, S. H., Elliott, C. G., Frantz, R. P., Hopper, R. K., Horn, E. M., Nicolls, M. R., Shlobin, O. A., Shah, S. J., Kovacs, G., Olschewski, H., and Rosenzweig, E. B., 2021, "Pulmonary Arterial Hypertension: Diagnosis, Treatment, and Novel Advances," Am J Respir Crit Care Med, 203(12), pp. 1472-1487.

[3] Humbert, M., McLaughlin, V., Gibbs, J. S. R., Gomberg-Maitland, M., Hoeper, M. M., Preston, I. R., Souza, R., Waxman, A., Escribano Subias, P., Feldman, J., Meyer, G., Montani, D., Olsson, K. M., Manimaran, S., Barnes, J., Linde, P. G., de Oliveira Pena, J., and Badesch, D. B., 2021, "Sotatercept for the Treatment of Pulmonary Arterial Hypertension," N Engl J Med, 384(13), pp. 1204-1215.

[4] Galiè, N., Humbert, M., Vachiery, J.-L., Gibbs, S., Lang, I., Torbicki, A., Simonneau, G., Peacock, A., Vonk Noordegraaf, A., Beghetti, M., Ghofrani, A., Gomez Sanchez, M. A., Hansmann, G., Klepetko, W., Lancellotti, P., Matucci, M., McDonagh, T., Pierard, L. A., Trindade, P. T., Zompatori, M., and Hoeper, M., 2015, "2015 ESC/ERS Guidelines for the diagnosis and treatment of pulmonary hypertension," European Respiratory Journal, 46(4), pp. 903-975.

[5] Benza, R. L., Gomberg-Maitland, M., Elliott, C. G., Farber, H. W., Foreman, A. J., Frost, A. E., McGoon, M. D., Pasta, D. J., Selej, M., Burger, C. D., and Frantz, R. P., 2019, "Predicting Survival in Patients With Pulmonary Arterial Hypertension: The REVEAL Risk Score Calculator 2.0 and Comparison With ESC/ERS-Based Risk Assessment Strategies," Chest, 156(2), pp. 323-337.

[6] Benza, R. L., Kanwar, M. K., Raina, A., Scott, J. V., Zhao, C. L., Selej, M., Elliott, C. G., and Farber, H. W., 2021, "Development and Validation of an Abridged Version of the REVEAL 2.0 Risk Score Calculator, REVEAL Lite 2, for Use in Patients With Pulmonary Arterial Hypertension," Chest, 159(1), pp. 337-346.

[7] Sweatt, A. J., Hedlin, H. K., Balasubramanian, V., Hsi, A., Blum, L. K., Robinson, W. H., Haddad, F., Hickey, P. M., Condliffe, R., Lawrie, A., Nicolls, M. R., Rabinovitch, M., Khatri, P., and Zamanian, R. T., 2019, "Discovery of Distinct Immune Phenotypes Using Machine Learning in Pulmonary Arterial Hypertension," Circ Res, 124(6), pp. 904-919.

[8] Wilkins, M. R., 2021, "Personalized Medicine for Pulmonary Hypertension:: The Future Management of Pulmonary Hypertension Requires a New Taxonomy," Clin Chest Med, 42(1), pp. 207-216.

[9] Oldham, W. M., Hess, E., Waldo, S. W., Humbert, M., Choudhary, G., and Maron, B. A., 2021, "Integrating haemodynamics identifies an extreme pulmonary hypertension phenotype," European Respiratory Journal, 58(2), p. 2004625.

[10] Rhodes, C. J., Wharton, J., Swietlik, E. M., Harbaum, L., Girerd, B., Coghlan, J. G., Lordan, J., Church, C., Pepke-Zaba, J., Toshner, M., Wort, S. J., Kiely, D. G., Condliffe, R., Lawrie, A., Gräf, S., Montani, D., Boucly, A., Sitbon, O., Humbert, M., Howard, L. S., Morrell, N. W., and Wilkins, M. R., 2022, "Using the Plasma Proteome for Risk Stratifying Patients with Pulmonary Arterial Hypertension," Am J Respir Crit Care Med.

[11] Rhodes, C. J., Wharton, J., Ghataorhe, P., Watson, G., Girerd, B., Howard, L. S., Gibbs, J. S. R., Condliffe, R., Elliot, C. A., Kiely, D. G., Simonneau, G., Montani, D., Sitbon, O., Gall, H., Schermuly, R. T., Ghofrani, H. A., Lawrie, A., Humbert, M., and Wilkins, M. R., 2017,

"Plasma proteome analysis in patients with pulmonary arterial hypertension: an observational cohort study," Lancet Respir Med, 5(9), pp. 717-726.
[12] Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I., 2020, "From local explanations to global understanding with explainable AI for trees," Nature Machine Intelligence, 2(1), pp. 56-67.
[13] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E., 2011, "Scikit-learn: Machine Learning in Python," J Mach Learn Res, 12, pp. 2825-2830.
[14] Jiang, H., Deng, Y., Chen, H. S., Tao, L., Sha, Q., Chen, J., Tsai, C. J., and Zhang, S., 2004, "Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes," BMC Bioinformatics, 5, p. 81.
[15] Breiman, L., 2001, "Random Forests," Machine Learning, 45(1), pp. 5-32.
[16] Shapley, L., 1951, "Notes on the n-Person Game -- II: THe Value of an n-Person Game," The RAND Corporation.
[17] Lundberg, S. M., and Lee, S.-I., 2017, "A unified approach to interpreting model predictions," Proceedings of the 31st International Conference on Neural Information Processing Systems, Curran Associates Inc., Long Beach, California, USA, pp. 4768–4777.
[18] Díaz-Uriarte, R., and Alvarez de Andrés, S., 2006, "Gene selection and classification of microarray data using random forest," BMC Bioinformatics, 7(1), p. 3.
[19] Niculescu-Mizil, A., and Caruana, R., 2005, "Predicting Good Probabilities With Supervised Learning," Proceedings of the 22nd International Conference on Machine LearningGermany.
[20] Bauer, Y., de Bernard, S., Hickey, P., Ballard, K., Cruz, J., Cornelisse, P., Chadha-Boreham, H., Distler, O., Rosenberg, D., Doelberg, M., Roux, S., Nayler, O., and Lawrie, A., 2021, "Identifying early pulmonary arterial hypertension biomarkers in systemic sclerosis: machine learning on proteomics from the DETECT cohort," The European respiratory journal, 57(6).
[21] Kanwar, M. K., Gomberg-Maitland, M., Hoeper, M., Pausch, C., Pittow, D., Strange, G., Anderson, J. J., Zhao, C., Scott, J. V., Druzdzel, M. J., Kraisangka, J., Lohmueller, L., Antaki, J., and Benza, R. L., 2020, "Risk stratification in pulmonary arterial hypertension using Bayesian analysis," European Respiratory Journal, p. 2000008.
[22] Errington, N., Iremonger, J., Pickworth, J. A., Kariotis, S., Rhodes, C. J., Rothman, A. M., Condliffe, R., Elliot, C. A., Kiely, D. G., Howard, L. S., Wharton, J., Thompson, A. A. R., Morrell, N. W., Wilkins, M. R., Wang, D., and Lawrie, A., 2021, "A diagnostic miRNA signature for pulmonary arterial hypertension using a consensus machine learning approach," EBioMedicine, 69, p. 103444.
[23] Cracowski, J.-L., Chabot, F., Labarère, J., Faure, P., Degano, B., Schwebel, C., Chaouat, A., Reynaud-Gaubert, M., Cracowski, C., Sitbon, O., Yaici, A., Simonneau, G., and Humbert, M., 2014, "Proinflammatory cytokine levels are linked to death in pulmonary arterial hypertension," European Respiratory Journal, 43(3), pp. 915-917.
[24] Soon, E., Holmes, A. M., Treacy, C. M., Doughty, N. J., Southgate, L., Machado, R. D., Trembath, R. C., Jennings, S., Barker, L., Nicklin, P., Walker, C., Budd, D. C., Pepke-Zaba, J., and Morrell, N. W., 2010, "Elevated Levels of Inflammatory Cytokines Predict Survival in Idiopathic and Familial Pulmonary Arterial Hypertension," Circulation, 122(9), pp. 920-927.

[25] Rabinovitch, M., Guignabert, C., Humbert, M., and Nicolls, M. R., 2014, "Inflammation and immunity in the pathogenesis of pulmonary arterial hypertension," Circ Res, 115(1), pp. 165-175.
[26] Groth, A., Vrugt, B., Brock, M., Speich, R., Ulrich, S., and Huber, L. C., 2014, "Inflammatory cytokines in pulmonary hypertension," Respir Res, 15, p. 47.
[27] Berghausen, E. M., Feik, L., Zierden, M., Vantler, M., and Rosenkranz, S., 2019, "Key inflammatory pathways underlying vascular remodeling in pulmonary hypertension," Herz, 44(2), pp. 130-137.
[28] Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., Steyerberg, E. W., Topic Group 'Evaluating diagnostic, t., and prediction models' of the, S. i., 2019, "Calibration: the Achilles heel of predictive analytics," BMC Med, 17(1), p. 230.

FIGURES

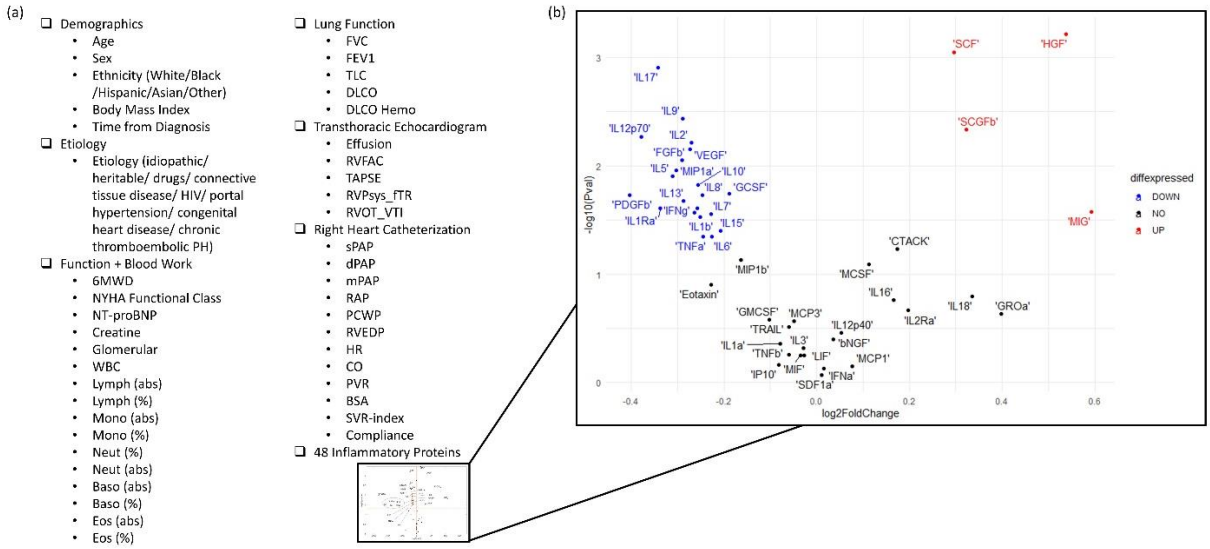

Figure 1. (a) An outline of all biomarkers considered; (b) Volcano plot showing all circulating proteins considered with fold change (FoldChange = concentration in high-risk patients/concentration in low-risk patients) in high 4-year risk, relative to low 4-year risk. Proteins in red and blue represent those that were statistically higher or lower (Pval < 0.05), respectively, and with a |FoldChange| > 1.

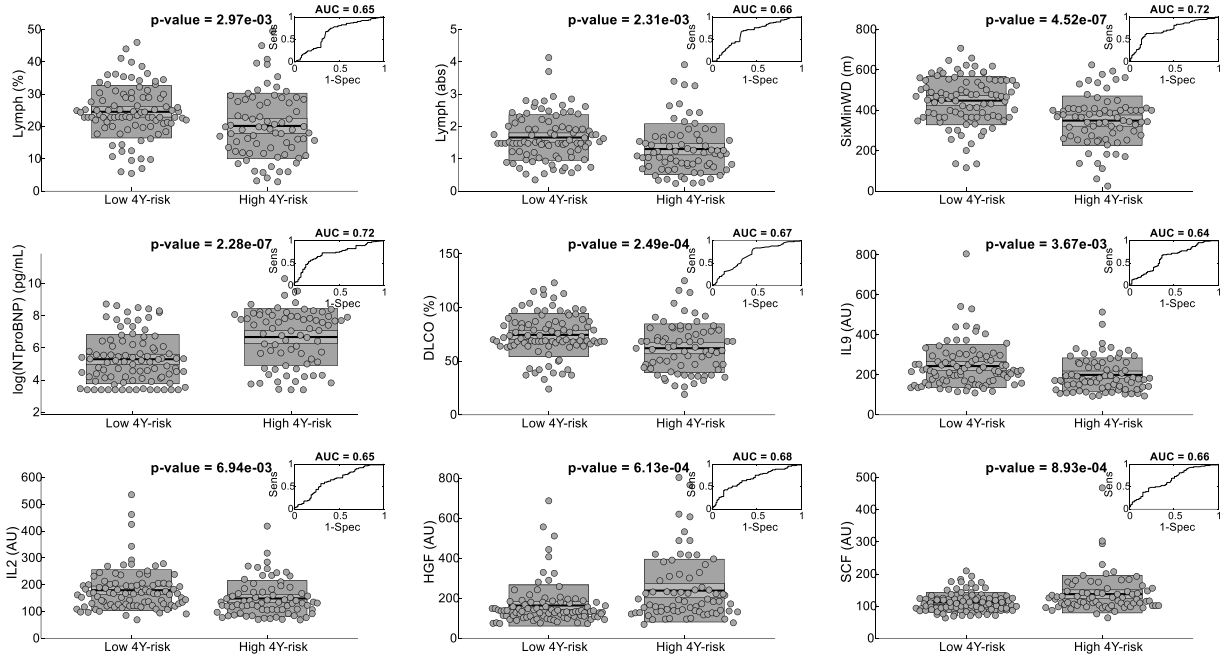

Figure 2. Mean comparison between low 4-year risk (4Y-risk) group and high 4Y-risk group for each of the 9 markers chosen through RFE. Each plot shows raw data, mean (red line), 95% CI in light shade, 1SD in dark shade. An inset in each plot shows the receiver operating curve for that marker along with the area under the curve (AUC). Note: Sens = sensitivity; Spec = specificity. H4Y-risk = High 4-year risk of death or need for transplant. L4Y-risk = low 4-year risk of death.

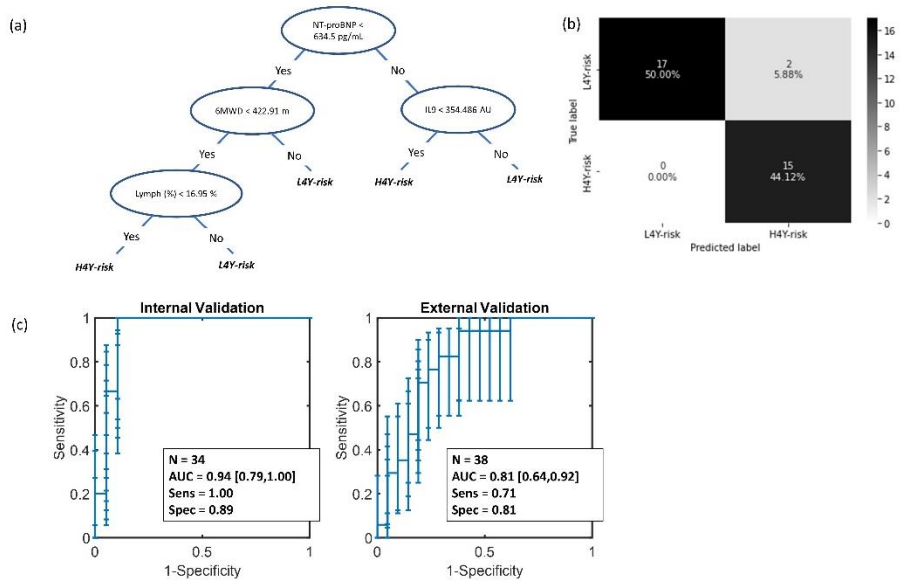

Figure 3. (a) An example decision tree randomly chosen from the 10,000 trees trained in the RF model. (b) Confusion matrix showing the accuracy of the RF model at predicting mortality in internal validation cohort. (c) Receiver operating curves (with error bars for each pointwise sensitivity calculation found by sampling 1000 bootstrap replicas) for the internal and external validation cohorts. AUC = area under receiver operating curve with 95% confidence interval in brackets. H4Y-risk = High 4-year risk of death or need for transplant. L4Y-risk = low 4-year risk of death.

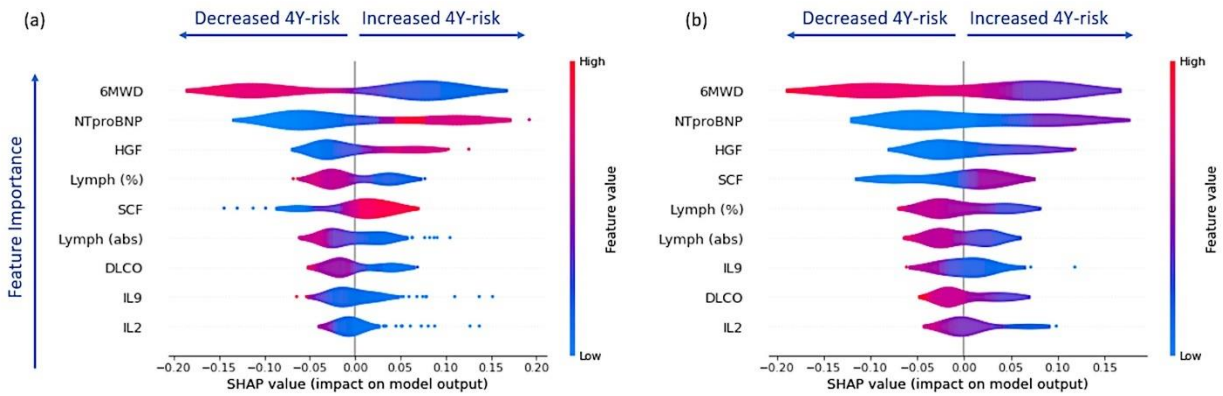

Figure 4. Global and Local Model Interpretability - Summary violin plot of SHAP values using the training dataset (a) and the internal validation dataset (b).

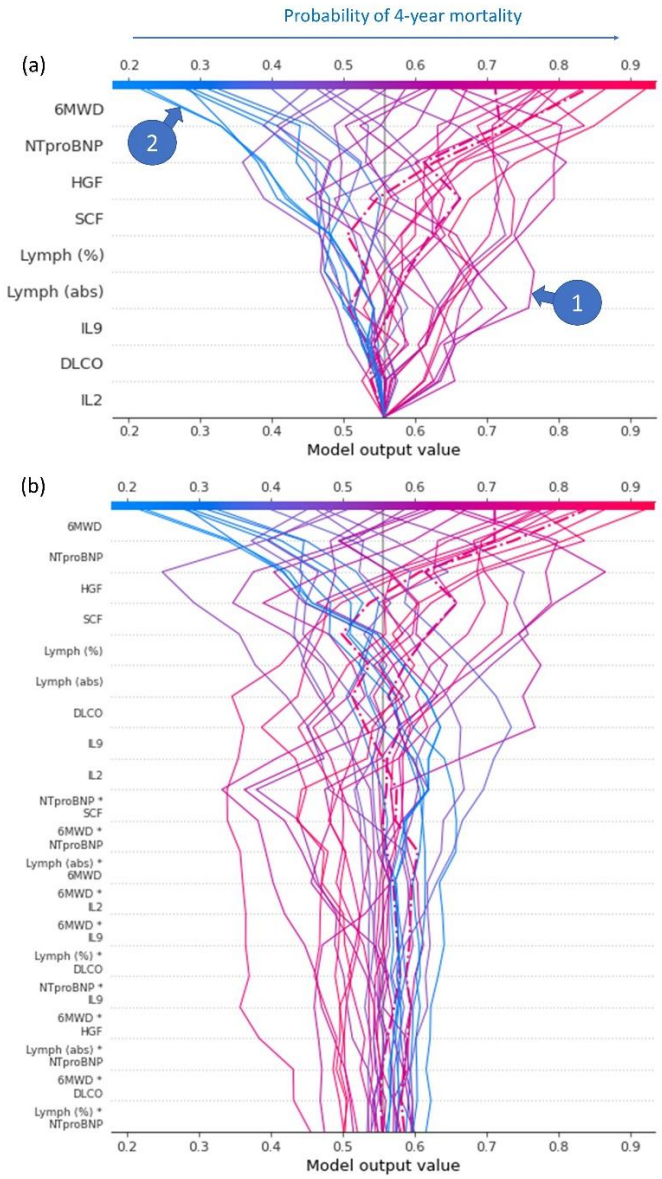

Figure 5. Decision plot of for the 34 patients withheld for (internal) validation. Each line corresponds to a given patient, and hidden lines correspond to the 2 misclassified patients. Figures (a) and (b) show decision plots with and without interactions, respectively.

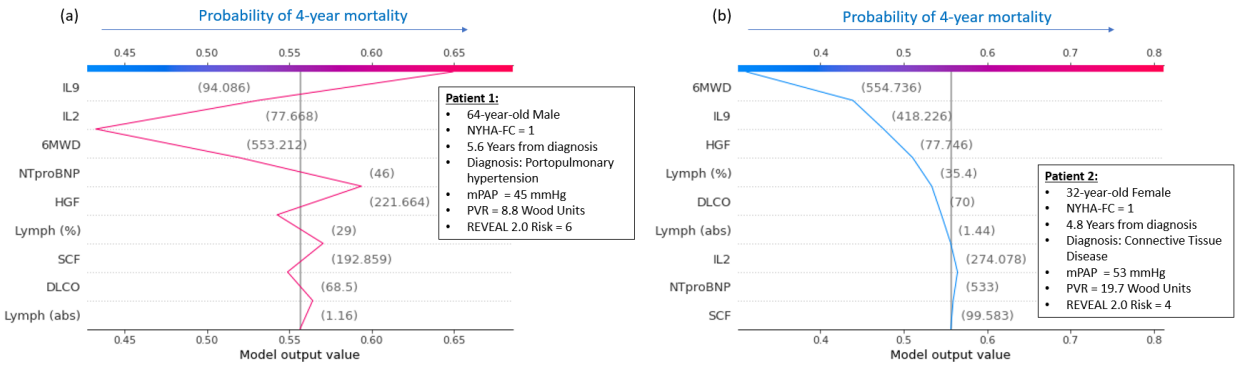

Figure 6. Decision plots for two randomly selected patients (a and b) show a graphical example of how a clinician can interpret the model prediction for a specific patient. The plot also shows that features ranked for an individual prediction can be notably different than the global model structure.

ONLINE SUPPLEMENTARY MATERIAL

O1. Expanded discussion of results

We were surprised to find that recursive feature elimination did not reveal IL6, IL10, tumor necrosis factor (TNF)-α, and/or IL1β as the most important variables for predicting 4Y-risk, given that previous studies reported them to be highly predictive of outcomes in PAH [1-4]. This is especially true for IL6, which is a pro-inflammatory cytokine believed to be one of the most important in the pathogenesis of PAH [5, 6]. One possible explanation for this finding is that many interleukin cytokines in our proteomic panel turned out to be highly correlated with each other (see Fig. E1), which interferes with the RF algorithm's ability to identify the strongest predictors [7]. In smaller datasets, like the one used in this study, this can be mitigated by RFE [7, 8]. Therefore, we allowed the RFE algorithm to choose the optimal 9 targets and avoid introducing bias into the outcome with pre-determined biomarkers. However, we wanted to know if replacing IL2 and IL9 with markers more prevalent in the literature would drastically change model performance. Fig. E2a shows a confusion matrix for the testing cohort if the model is trained after replacing IL2 and IL9 with IL6 and IL10, respectively. Based on the testing dataset, withheld from training, the model resulted in a sensitivity of 93.3% and specificity of 87.5% (F1 score = 90.3%). Global interpretability analysis shows that IL6 and IL10 are the least important predictors (see Fig. E2b). Fig. E2c shows a confusion matrix for the testing cohort if the model is trained after replacing IL2 and IL9 with IL6 and IL1β, respectively. Based on the testing dataset, the final model resulted in a sensitivity of 80.0% and specificity of 80.0% (F1 score = 80.0%). Global model structure analysis again showed these cytokines to be the least important predictors (see Fig. E2d). Interestingly, replacing IL1β with TNF-α produced identical results (results not shown).

O2. Model validation using an external dataset

Thirty-eight (N = 38, see Table E1) patients evaluated at Sheffield University between 2008 and 2014 were analyzed as an external validation cohort. All proteomic biomarkers were analyzed at Stanford University from shipped blood samples, but all other data was directly measured at Sheffield University. Several inconsistencies in data collected prevented this dataset from serving as a direct comparison:

(1) 6MWD was estimated as: $e6MWD = 0.7225 \cdot MSWT + 70.769$, based on a relationship from ref. [9]. *To arrive at this equation, we downloaded the figure showing the correlation ($r = 0.863$, $p < 0.0001$) in 44 patients with chronic obstructive pulmonary disease (COPD). The original figure included a regression line, which was sampled at 4 random coordinate points using the WebPlotDigitizer [10]. We then fit our own regression line to those 4 coordinates to arrive at the estimate of 6MWD (e6MWD) from MSWT.*

We note that there are multiple limitations to utilizing this relationship in our study. Firstly, it was derived from a modest cohort of COPD patients and utilized for PAH patients. Second, while the correlation is statistically significant, the error between the two measurements is likely high. Although the raw data was not available to perform Bland-Altman analysis, this can be concluded from qualitatively inspecting the published correlation image.

(2) NT-proBNP was measured in the plasma of the training cohort, but in the serum of the external validation cohort. There have been mixed findings on the comparison between NT-proBNP measurements in plasma and serum, which can also be impacted by the assay platform [11, 12]. This speaks to the potential inconsistencies that can arise between different institutions

across the globe, which highlights the need for either training all clinically implemented datasets on extremely large cohorts or standardizing the biochemical analysis.

Fig. E3 shows the 9 features compared in the Stanford and Sheffield cohorts. The trends between low risk and high-risk groups are consistent between the two cohorts, but there is also a bias in multiple markers. Therefore, the final RF algorithm (trained on 100% of the Stanford cohort) was trained and tested (on 100% of the Sheffield cohort) based on measurements relative to the low-risk groups to standardize the thresholds in the decision branches.

Fig. 3c shows a receiver operating curve of the H4Y-risk probability predicted for the Sheffield validation cohort alongside with the internal validation cohort. In spite of the aforementioned difference in data acquisition between the Stanford and the Sheffield cohorts, the RF model (trained on the Stanford cohort) is able to predict 4Y-risk with 71% sensitivity and 81% specificity in the Sheffield external validation cohort (AUC = 0.81, 95% CI = 0.64-0.92).

Global interpretability analysis (see Fig. E4a) of the Sheffield cohort revealed that, even though 6MWD was estimated from MSWT (e6MWD), it is still the most important feature for predicting 4-year risk. Furthermore, NT-proBNP was also one of the top predictors. However, unlike in the Stanford cohort, lung function (measured by DLCO) was significantly reduced in the Sheffield cohort (see Fig. E3) and heavily influenced the prediction score.

Decision plots of all the patients in the Sheffield cohort (see Fig. E4b) revealed that the majority of misclassified patients tended to fall within a probability score that was closer to 50% (near the baseline predicted score before any features are considered) than those patients correctly identified as high or low 4-year risk.

O3. Comparison against the REVEAL 2.0 risk calculator

Although it is not a fair comparison - because the REVEAL 2.0 risk calculator was developed to predict 1-year risk – we applied the REVEAL 2.0 calculator to predicting 4-year risk in an attempt to compare our RF model performance against a known standard. Fig. E5 shows receiver operating curves for the probability estimates of the RF model, applied to the internal and external validation cohorts (Fig. E5a and E5b), alongside the REVEAL 2.0 scores (Fig. E5c). Pointwise confidence intervals on the sensitivity calculation were computed using vertical averaging from 1000 sampled bootstrap replicas.

Applying the REVEAL 2.0 calculator to the entire cohort considered in this study (N = 167) revealed that a score above 9 can predict 4-year risk with 54% sensitivity and 86% specificity. Based on the area under the curve (AUC), the probability of a patient having a poor 4-year outcome, given that the REVEAL 2.0 calculator predicted that patient would have a poor outcome, is 78% (with a 95% confidence interval of 70-84%). This result is consistent with previous studies [13] and shows that the REVEAL 2.0 calculator is an effective estimate beyond the 1-year window.

Although larger studies are needed, Fig. E5 shows that the performance of the RF model trained in this study is comparable to the REVEAL 2.0 calculator. Therefore, the computational pipeline proposed in this study offers a reliable prediction algorithm that is interpretable, and well suited for machine-physician interaction in patient-tailored therapy.

TABLES

	H4Y-risk	L4Y-risk	$P_{2\text{-tailed}}$
Sample Size	24	14	
Age (years ± SD)	60.0 ± 8.95	52.9 ± 15.4	0.080
% Female Patients	54.1	78.5	
Race			
White	24	12	0.057
Asian	0	2	
Hispanic	0	0	
Black	0	0	
Other	0	0	
PAH subtype			
Connective Tissue Disease	9	5	0.401
Idiopathic PAH	11	7	
Drug and Toxins	0	0	
Congenital Heart Disease	1	2	
Portopulmonary Hypertension	3	0	
Hereditary PAH	0	0	
NYHA Functional Class*			
Class I	0	0	0.022
Class II	1	5	
Class III	22	8	
Class IV	1	0	
Hemodynamics			
mPAP (mmHg)	45.88 ± 11.41	52.57 ± 16.84	0.153
PVR (dyn·s/cm^5)	7.20 ± 3.52	10.58 ± 5.80	0.031
Cardiac Index (m^2·min)	3.08 ± 0.69	2.50 ± 0.70	0.018
Mean Right Atrial Pressure (mmHg)	11.58 ± 7.08	10.79 ± 6.09	0.729
PCWP (mmHg)	9.80 ± 3.71	12.0 ± 5.39	0.145

Table E1. Sheffield cohort patient characteristics in H4Y-risk and L4Y-risk groups. Continuous data is compared using a t-test and categorical variables by χ^2 test. *Note: NYHA Functional Class was not available for one patient.

FIGURES

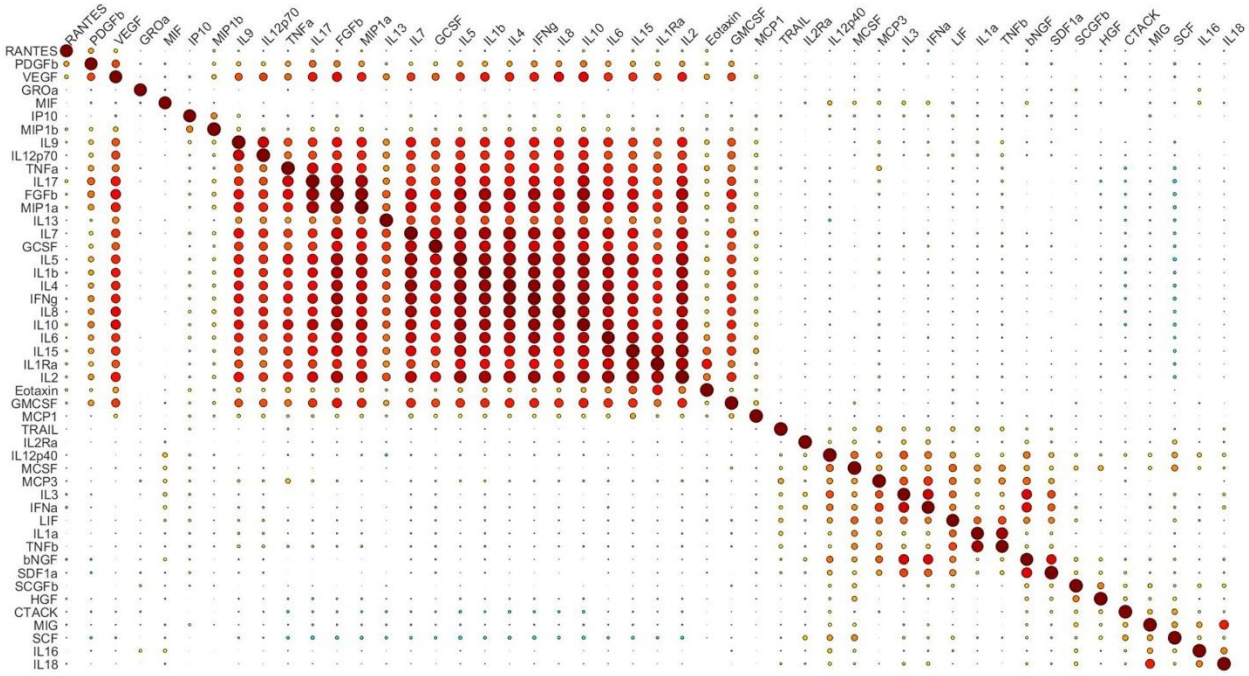

Figure E1. Correlation matrix between markers in the proteomic inflammatory panel. Each correlation is represented by a coloured circle. Both the circle size and colour represent the strength of the correlation. The diagonal points, corresponding to a correlation coefficient: R^2 = 1, can be used as reference.

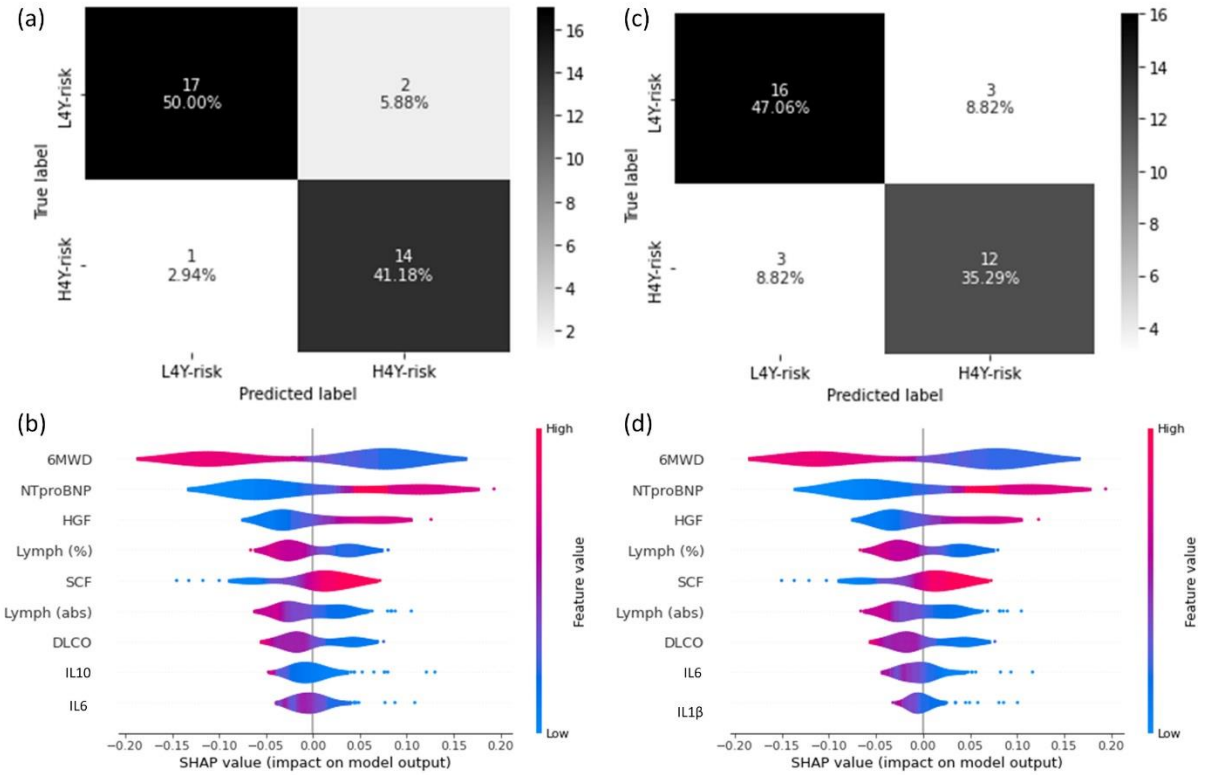

Figure E2. Confusion matrices and global interpretability analysis generated with the internal validation dataset when considering: (a) the 9 features chosen by the original RFE algorithm, but with IL2 and IL9 replaced with IL6 and IL10, respectively, and (b) and global model structure corresponding to the confusion matrix in (a); (c) the 9 features chosen by the original RFE algorithm, but with IL2 and IL9 replaced with IL6 and IL1β, respectively, and (d) global model structure corresponding to the confusion matrix in (c).

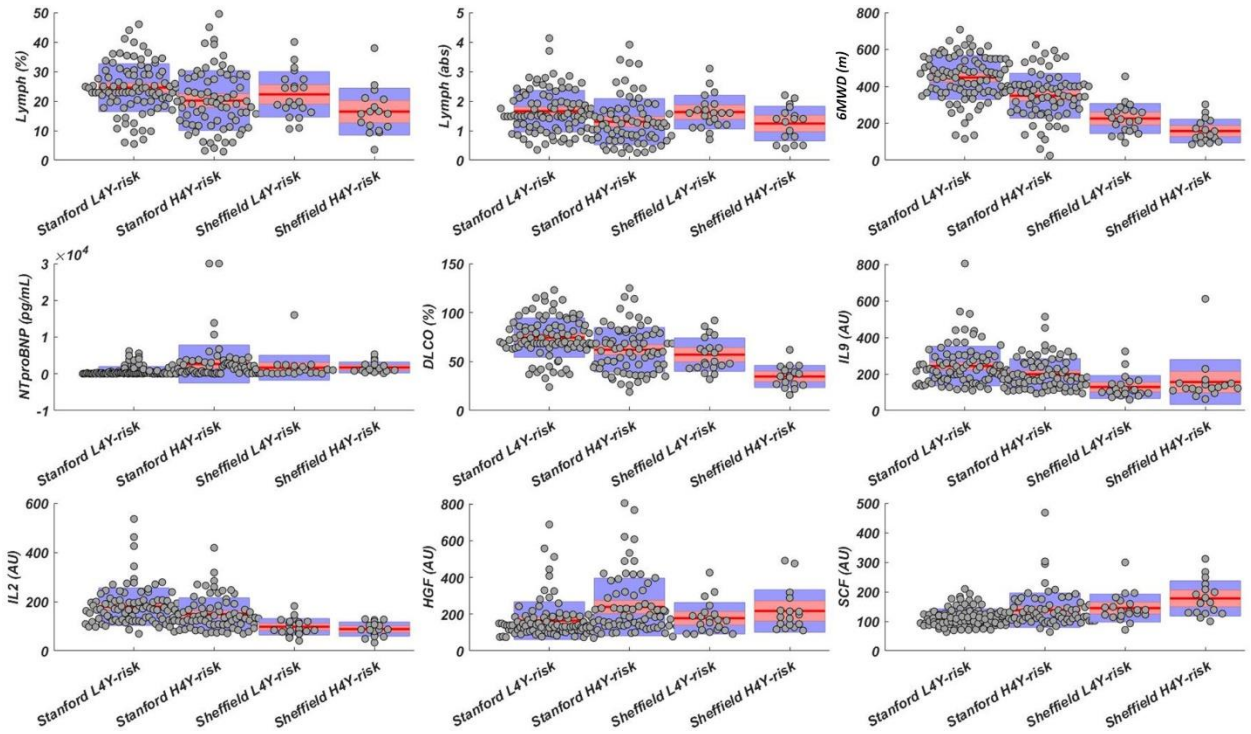

Figure E3. Mean comparison between the low 4-year risk (L4Y-risk) group and high 4-year risk (H4Y-risk) groups within the Stanford and Sheffield Cohorts. Note: estimated 6MWD and serum NT-proBNP levels are shown for the Sheffield cohort.

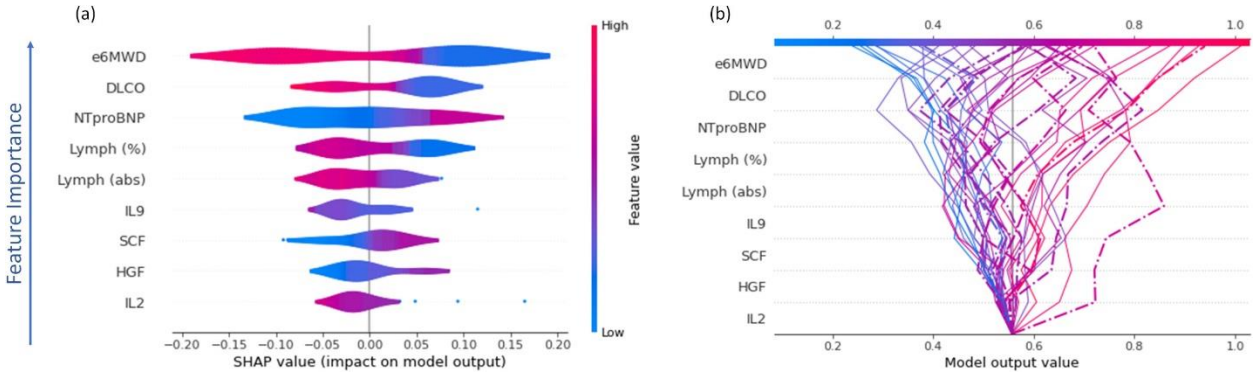

Figure E4. Violin plot (a) and decision plot (b) for global interpretability analysis of the external validation cohort. Dashed lines in the decision plots represent patients that were incorrectly classified. Note: e6MWD is the estimated 6MWD from MSWT.

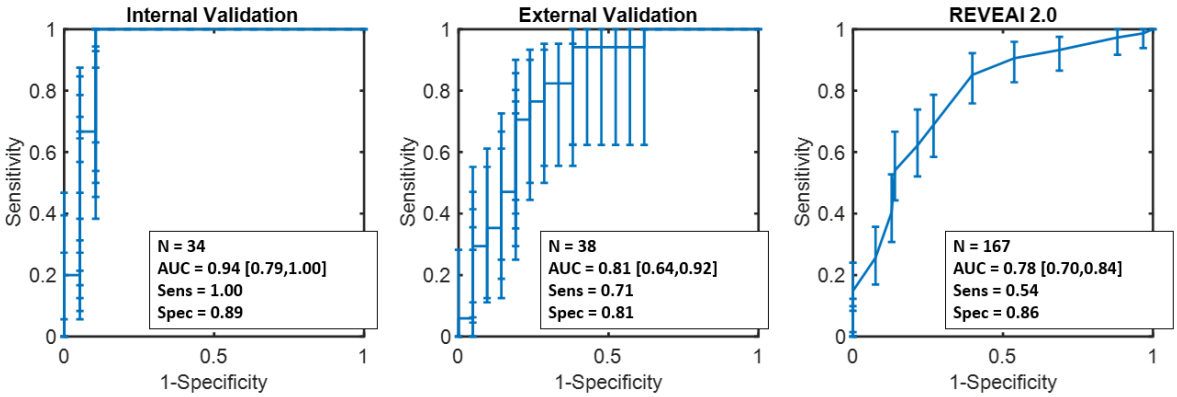

Figure E5. (a and b) Receiver operating curves of the probability estimates of the RF model for the internal and external validation cohorts. (c) ROC for the REVEAL 2.0 risk score. Each plot shows the area under the curve (AUC) with the 95% confidence interval in brackets. Note: N = sample size used to generate the curve; Sens = sensitivity; Spec = specificity. Error bars for pointwise sensitivity calculations were found by sampling 1000 bootstrap replicas.

REFERENCES FOR ONLINE SUPPLEMENTARY MATERIAL

[1] Cracowski, J. L., Chabot, F., Labarère, J., Faure, P., Degano, B., Schwebel, C., Chaouat, A., Reynaud-Gaubert, M., Cracowski, C., Sitbon, O., Yaici, A., Simonneau, G., and Humbert, M., 2014, "Proinflammatory cytokine levels are linked to death in pulmonary arterial hypertension," The European respiratory journal, 43(3), pp. 915-917.

[2] Heresi, G. A., Aytekin, M., Hammel, J. P., Wang, S., Chatterjee, S., and Dweik, R. A., 2014, "Plasma interleukin-6 adds prognostic information in pulmonary arterial hypertension," The European respiratory journal, 43(3), pp. 912-914.

[3] Rabinovitch, M., Guignabert, C., Humbert, M., and Nicolls, M. R., 2014, "Inflammation and immunity in the pathogenesis of pulmonary arterial hypertension," Circ Res, 115(1), pp. 165-175.

[4] Soon, E., Holmes, A. M., Treacy, C. M., Doughty, N. J., Southgate, L., Machado, R. D., Trembath, R. C., Jennings, S., Barker, L., Nicklin, P., Walker, C., Budd, D. C., Pepke-Zaba, J., and Morrell, N. W., 2010, "Elevated levels of inflammatory cytokines predict survival in idiopathic and familial pulmonary arterial hypertension," Circulation, 122(9), pp. 920-927.

[5] Groth, A., Vrugt, B., Brock, M., Speich, R., Ulrich, S., and Huber, L. C., 2014, "Inflammatory cytokines in pulmonary hypertension," Respir Res, 15, p. 47.

[6] Berghausen, E. M., Feik, L., Zierden, M., Vantler, M., and Rosenkranz, S., 2019, "Key inflammatory pathways underlying vascular remodeling in pulmonary hypertension," Herz, 44(2), pp. 130-137.

[7] Darst, B. F., Malecki, K. C., and Engelman, C. D., 2018, "Using recursive feature elimination in random forest to account for correlated variables in high dimensional data," BMC Genetics, 19(1), p. 65.

[8] Gregorutti, B., Michel, B., and Saint-Pierre, P., 2017, "Correlation and variable importance in random forests," Statistics and Computing, 27(3), pp. 659-678.

[9] Gopalakrishnan, R., and Singla, R., 2018, "Correlation of a modified shuttle walk with six-minute walk test in COPD patients," European Respiratory Journal, 52(suppl 62), p. PA2451.

[10] Drevon, D., Fursa, S. R., and Malcolm, A. L., 2017, "Intercoder Reliability and Validity of WebPlotDigitizer in Extracting Graphed Data," Behav Modif, 41(2), pp. 323-339.
[11] Peake, R. W., Turner, H. E., Leaper, W., Deans, K. A., Hannah, A., and Croal, B. L., 2012, "Comparison of sample types for N-terminal pro-B-type natriuretic peptide measured on the Siemens Immulite 2500 and Dimension Vista LOCI methods," Ann Clin Biochem, 49(Pt 5), pp. 494-496.
[12] Lippi, G., Salvagno, G. L., Montagnana, M., and Guidi, G. C., 2007, "Measurement of Elecsys NT-proBNP in serum, K2 EDTA and heparin plasma," Clin Biochem, 40(9-10), pp. 747-748.
[13] Benza, R. L., Gomberg-Maitland, M., Elliott, C. G., Farber, H. W., Foreman, A. J., Frost, A. E., McGoon, M. D., Pasta, D. J., Selej, M., Burger, C. D., and Frantz, R. P., 2019, "Predicting Survival in Patients With Pulmonary Arterial Hypertension: The REVEAL Risk Score Calculator 2.0 and Comparison With ESC/ERS-Based Risk Assessment Strategies," Chest, 156(2), pp. 323-337.