Machine learning to differentiate pulmonary hypertension due to left heart disease from pulmonary arterial hypertension

Katleen Swinnen[1,8], Kenneth Verstraete [1,2,8], Claudia Baratto[3], Laura Hardy [1], Maarten De Vos[2,4], Marko Topalovic[5], Guido Claessen [6], Rozenn Quarck [7], Catharina Belge[7], Jean-Luc Vachiery[3], Wim Janssens [7] and Marion Delcroix [7]

[1]Laboratory of Respiratory Diseases and Thoracic Surgery (BREATHE), Department of Chronic Diseases and Metabolism (CHROMETA), KU Leuven, Leuven, Belgium. [2]STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, Department of Electrical Engineering (ESAT), KU Leuven, Leuven, Belgium. [3]Department of Cardiology, Ospedale San Luca IRCCS Istituto Auxologico Italiano, Milan, Italy. [4]Department of Development and Regeneration, KU Leuven, Leuven, Belgium. [5]ArtiQ NV, Leuven, Belgium. [6]Department of Cardiology, University Hospitals of Leuven, Leuven, Belgium. [7]Clinical Department of Respiratory Diseases, Centre of Pulmonary Vascular Diseases, University Hospitals of Leuven, BREATHE, CHROMETA, KU Leuven, Leuven, Belgium. [8]These authors contributed equally.

Corresponding author: Marion Delcroix (marion.delcroix@uzleuven.be)

Abstract

Background and aims Pulmonary hypertension due to left heart disease (PH-LHD) is the most frequent form of PH. As differential diagnosis with pulmonary arterial hypertension (PAH) has therapeutic implications, it is important to accurately and noninvasively differentiate PH-LHD from PAH before referral to PH centres. The aim was to develop and validate a machine learning (ML) model to improve prediction of PH-LHD in a population of PAH and PH-LHD patients.

Methods Noninvasive PH-LHD predictors from 172 PAH and 172 PH-LHD patients from the PH centre database at the University Hospitals of Leuven (Leuven, Belgium) were used to develop an ML model. The Jacobs score was used as performance benchmark. The dataset was split into a training and test set (70:30) and the best model was selected after 10-fold cross-validation on the training dataset (n=240). The final model was externally validated using 165 patients (91 PAH, 74 PH-LHD) from Erasme Hospital (Brussels, Belgium).

Results In the internal test dataset (n=104), a random forest-based model correctly diagnosed 70% of PH-LHD patients (sensitivity: n=35/50), with 100% positive predicted value, 78% negative predicted value and 100% specificity. The model outperformed the Jacobs score, which identified 18% (n=9/50) of the patients with PH-LHD without false positives. In external validation, the model had 64% sensitivity at 100% specificity, while the Jacobs score had a sensitivity of 3% for no false positives.

Conclusions ML significantly improves the sensitivity of PH-LHD prediction at 100% specificity. Such a model may substantially reduce the number of patients referred for invasive diagnostics without missing PAH diagnoses.

Introduction

Pulmonary arterial hypertension (PAH) has been defined haemodynamically by pre-capillary pulmonary hypertension (PH), with a mean pulmonary arterial pressure (mPAP) $\geqslant$20 mmHg and a pulmonary arterial wedge pressure (PAWP) $\leqslant$15 mmHg, measured at right heart catheterisation (RHC) [1]. If the PAWP is higher than 15 mmHg, post-capillary PH is present and pulmonary hypertension due to left heart disease (PH-LHD) should be considered. By contrast to PAH, PH-LHD is frequent and accounts for 60–80% of all PH patients [2].

Obtaining an accurate diagnosis of PAH can be challenging considering that PAH is relatively rare, symptoms are nonspecific and potentially related to other cardio-pulmonary diseases. Moreover, identifying the correct cause of PH is even more difficult in an ageing population with comorbidities. Proper distinction between PAH and PH-LHD may be difficult, especially for idiopathic PAH (IPAH) and PH due to heart failure with preserved ejection fraction (HFpEF), a common cause of PH-LHD. The final differential diagnosis between pre- and post-capillary PH, *i.e.,* between PAH and PH-LHD, relies only on PAWP being lower or higher than 15 mmHg.

In practice, measuring and correctly interpreting PAWP can be difficult and must therefore be performed in experienced centres. A correct differential diagnosis is important considering that the use of PAH-approved therapies is not recommended in PH-LHD and can even be harmful [2]. Over the past years, increased PH awareness and phenotype overlap of patients with PH-LHD and PAH have led to a higher number of referrals of suspected PAH patients to PH centres, resulting in a significantly increased number of invasive RHC procedures [2].

Considering the limited though potential risk of complications, invasiveness and costs of RHC procedures, various prediction models to differentiate PAH from PH-LHD were developed to optimise the referral of PAH patients [3–8]. However, currently existing prediction models are limited by restricted predictive value, low number of predictors, inclusion of invasive predictors and lack of external validation. In addition, these models are developed using logistic regression and are therefore limited by the assumption that each predictor is linearly related to PH-LHD diagnosis, while practice indicates that more complex relationships should be considered, although still unexplored [9–11].

We hypothesised that a machine learning (ML) prediction model based on noninvasive clinically available variables would be a promising alternative and innovative method for the differential diagnosis of PAH and PH-LHD, accounting for more complex interactions than currently existing logistic regression models. Since PAH is a life-threatening disease with a mean survival of only 2.8 years if left untreated, a false positive result implies that the model classifies the patient as having PH-LHD, overlooking PAH and preventing the patient from benefitting from efficient PAH therapies.

We therefore aimed to develop a more sensitive and highly specific prediction ML model for PH-LHD on noninvasive parameters. Such a model could reduce the number of RHC in patients with a high pre-test likelihood of PH-LHD and restrict RHC to a group with a higher pre-test likelihood of PAH.

Methods
Patient population
The target population consisted of all consecutive idiopathic, heritable and drug- or toxin-induced PAH and PH-LHD patients diagnosed in alignment with the haemodynamic definition of the 2015 ESC/ERS guidelines for the diagnosis and treatment of pulmonary hypertension at the University Hospitals of Leuven (Leuven, Belgium) and documented within the PH database between January 2000 and March 2020 (supplementary figure S1). All included patients underwent invasive haemodynamic diagnosis in accordance with the 2015 ESC/ERS guidelines for PH [12]. A complete echocardiographic examination was performed [13]. The clinical diagnosis of PAH or PH-LHD was based on standard of care tests and a final evaluation by the local PH team, resulting in a study cohort of 172 PAH and 172 PH-LHD patients. If a PAWP was unmeasurable, we used left ventricle end diastolic pressure to diagnose PAH *versus* PH-LHD. The study was approved by the Ethical Committee of University Hospitals of Leuven. This cohort was used to develop and internally validate our prediction models.

Development of the prediction models
To develop the prediction models, a list of potential PH-LHD predictors was compiled after a literature review. Parameters used in other predictive scores for HFpEF or PH-LHD were included [3–8]. The list was adjusted based on the availability of the data, excluding parameters with a high percentage of missing data or low availability in routine practice. Afterwards the list was reviewed independently by three experts (two PH experts and one heart failure expert). The final list comprised demographics, medical history, echocardiography, lung function tests, laboratory blood parameters and ECG variables (full list in table 1). The predictors were recorded manually from electronic medical files from the University Hospitals of Leuven. The final database comprised 64 500 values. Variables with >60% missing values were excluded from the analysis. The dataset was randomly split (70:30) in a training dataset (n=240), for model development and model selection, and an independent test dataset (n=104), for internal validation of the selected model. After internal validation in the test set, the model was retrained on the entire Leuven cohort (n=344) before external validation on the second cohort of the Erasme Hospital in Brussels (Belgium).

TABLE 1 Characteristics of the University Hospitals of Leuven cohort

Clinical characteristics	PAH	PH-LHD	p-value
Patients n	172	172	
Age at diagnosis years, median (IQR)	54 (22)	67 (12)	<0.001
Male	63 (37)	52 (30)	0.20
BMI kg·m^{-2}	27±6	30±7	<0.001
NYHA class			
I	8 (5)	14 (8)	0.20
II	44 (26)	56 (33)	
III	91 (53)	81 (47)	
IV	28 (16)	21 (12)	
Medical history			
Diabetes mellitus type II	50 (29)	72 (42)	<0.001
Hypertension	76 (44)	112 (65)	<0.01
Hypercholesterolaemia	53 (31)	79 (46)	0.03
Obesity	52 (30)	83 (48)	<0.001
Valvular surgery without residual valve disease	4 (2)	12 (7)	<0.001
Left heart disease	15 (9)	32 (19)	<0.001
Smoking			
Current smoker	8 (5)	7 (4)	0.65
Former smoker	52 (30)	60 (35)	
Never-smoker	112 (65)	105 (61)	
ECG parameters			
Atrial fibrillation	45 (26)	114 (66)	<0.001
SV1+RV6 mm	13±5	16±8	<0.001
R axis degrees	80±56	31±60	<0.001
PR interval ms	168±31	175±35	0.02
PRT axes degrees	54 ±21	51±20	0.10
QRS duration ms	98±17	103±28	0.02
Echocardiography parameters			
Left atrial dilation worse than mild	16 (9)	59 (34)	<0.001
Left valvular disease worse than mild	5 (3)	44 (26)	<0.001
Left ventricle end diastolic diameter mm	37.8±7.1	44.3±7.8	<0.001
Mitral valve disease worse than mild	2 (1)	38 (22)	<0.001
Mitral E/A ratio	1.0±0.5	1.9±0.9	<0.001
Mitral E V'_{max} mm·s^{-1}	66±25	104±41	<0.001
Mitral E/E′ ratio	9.6±5.7	10.7±9.7	0.03
Dilation of the right ventricle			
Not dilated (qualitative)	26 (15)	46 (27)	<0.001
Mildly dilated (qualitative)	43 (25)	28 (16)	
Severely dilated (qualitative)	57 (33)	21 (12)	
Pulmonary function test parameters			
FVC L	3±3.1	2.6±2.5	0.05
FEV$_1$ L	2.1±0.7	2.0±0.7	<0.001
FEV$_1$/FVC ratio	74±11	75 ±8	0.44
Peak expiratory flow L·s^{-1}	6±5	6±2	0.11
D_{LCO}	4.4±1.7	4.4±1.8	0.69
Laboratory blood tests			
Haemoglobin g·dL^{-1}	16±2	14±1	0.87
Haematocrit %	43±9	37±10	0.49
Haemodynamic parameters			
Right atrial pressure mmHg	8±5	14±6	<0.001
Systolic PAP mmHg	78±20	66±19	0.66
Diastolic PAP mmHg	32±10	27±8	0.01
Mean PAP mmHg	49±12	41±11	<0.001
CO L·min^{-1}	3.7±1.2	4.9±1.7	<0.001
CI L·min^{-1}·m^{-2}	2.11±0.68	2.59±0.85	<0.001
S_{vO_2}%	61±10	65±9	0.59
PVR Wood Units	12±26	5±3	<0.001
PVR dynes·s·cm^{-5}	958±504	377±280	<0.001
PAWP mmHg	9±3	22±6	<0.001

Values are expressed as n (%) or mean±sd unless indicated otherwise. PAH: pulmonary arterial hypertension; PH-LHD: pulmonary hypertension due to left heart disease; IQR: interquartile range; BMI: body mass index; NYHA class: New York Heart Association functional class; SV1+RV6: sum of the S deflection in V1 and R deflection in V6 in millimetres on ECG; Mitral E/A: the peak early diastolic (E), atrial systolic (A) transmitral flow velocities for evaluation of left ventricle diastolic function; Mitral E V'_{max}: peak E mitral flow velocity; Mitral E/E′: ratio of early diastolic mitral inflow velocity to early diastolic mitral annulus velocity for the evaluation of left ventricular filling pressure; FVC: forced vital capacity; FEV$_1$: forced expiratory volume in 1 s; FEV$_1$/FVC: Tiffeneau index; D_{LCO}: diffusing capacity of the lung for carbon monoxide; PAP: pulmonary arterial pressure; CO: cardiac output; CI: cardiac index; S_{vO_2}: mixed venous oxygen saturation; PVR: pulmonary vascular resistance; PAWP: pulmonary arterial wedge pressure.

Logistic regression model

We first developed a logistic regression model, in which the effect of each predictor variable was evaluated by univariate logistic regression in the training dataset. Variables with a p-value <0.10 were imputed in a multivariable logistic regression model with stepwise forward method, which determined the final model. The performance of the model was studied according to clinical performance and compared to the Jacobs score in the independent test dataset.

Machine learning model

To develop an ML model, we explored different classification models (k-nearest neighbours, support vector machine, random forest and gradient boosting). For each model, the hyperparameters were optimised using grid search and 10-fold cross-validation [14]. The latter allowed us to obtain a better estimate of generalisation performance by iterating over 10 different folds of the training set. To reduce the risk of false positives, this entire procedure was repeated using bootstrapping (random sampling with replacement [15], n=240) with 100 samples [16] of the training set, resulting in 100 trained models for each algorithm (supplementary figure 2). In a voting fashion, a subject was then classified as PH-LHD if and only if at least 95% of the bootstrap models predicted a probability of at least 0.5; allowing for 5% outlier models that might have been trained on extreme variations of the training set during bootstrapping. We used sensitivity as optimisation metric, excluding models with positive predicted value (PPV) <100% in the training set. The algorithm that performed best over this entire procedure (highest sum of sensitivity and 100% specificity) was selected for validation on the independent test set. For the selected algorithm, we analysed its performance in function of the number of parameters to reduce the number of required parameters in clinical practice. We selected the features based on feature importance provided by the model and domain knowledge by the PH specialists, and selected the number based on the results during a new iteration of 10-fold cross-validation. The Shapley additive explanations (SHAP) methodology [17] was used to explain the predictions of the ML model on an individual level and facilitate the interpretation of model results.

The Jacobs score as benchmark

The Jacobs score, an existing risk score based on a logistic regression model, was used as benchmark to compare performance of the prediction models [4]. The Jacobs score includes four predictors: left heart disease in medical history (yes/no), sum of the S wave deflection in V1 and R wave deflection in V6 on the electrocardiogram (ECG), left atrial dilation on transthoracic echocardiography (TTE) (yes/no) and left ventricle valve disease worse than mild on TTE (yes/no) [4]. As suggested by the authors, risk score cut-off ⩾72 (range of the score: 0–96 points) was used to identify patients with PH-LHD [4].

First, as means of external validation for the Jacobs score itself, the performance of the score was validated in the entire retrospective Leuven cohort (n=344). We evaluated the clinical performance based on area under the receiver operating curve (AUC-ROC), sensitivity, specificity, PPV and negative predicted value (NPV) for the prediction of PH-LHD. Secondly, the performance of the logistic regression model and the ML model was compared to the Jacobs score in the internal independent test dataset (n=104). The best of the two models was then externally validated on the external validation dataset of Brussels (n=165) and compared to the Jacobs score.

External validation population

To determine the reproducibility and generalisability of the Jacobs score and our model to new and different patients, data from a cohort from Erasme Hospital Brussels were collected to externally validate the final model. All consecutive patients with an invasive diagnosis from 2000 onwards were eligible for inclusion. The same parameters as the subjects from University Hospitals of Leuven were manually recorded. A final diagnosis was based on the international PH guidelines, invasive haemodynamics and clinical interpretation of >two PH specialists at Erasme Hospital Brussels. Characteristics were compared between two cohorts.

Statistical analysis

The continuous variables are presented as mean±$_{SD}$ or median with interquartile range, while categorical variables are shown as counts and percentages. The Shapiro–Wilk test was used to test for normality. If significant, t-test was used to compare means, otherwise we applied the Mann–Whitney U test. Fisher's exact test was used to test differences in proportions for dichotomous variables and Chi-square test for categorical variables. Missing data were handled by multivariate imputation by chained equations (MICE) [18], during cross-validation. Bayesian ridge regression was used as estimator in the MICE method for imputing missing values. Variables of which the missingness was systematically related to unobserved data were excluded from the analysis. Categorical variables were encoded using one-hot encoding. The statistical

analyses were performed using SPSS 25.0 (IBM) or Scipy and Statsmodels for Python 3.8. Python was also used for the ML model. All p-values were two-sided and significance level was set at 0.05.

Results

Study population

Patient characteristics for the University Hospitals of Leuven cohort are presented in table 1. The median age of PH-LHD patients (n=172) was 67 years and significantly higher than that of PAH patients (n=172; 54; p<0.001). Haemodynamic parameters including right atrial pressure, mPAP, PAWP, cardiac index and pulmonary vascular resistance were significantly different between the two groups. The median body mass index (BMI) was 27±6 for PAH and 30±7 for PH-LHD. A medical history of diabetes, hypertension, obesity, valvular surgery without residual valvular disease and left heart disease were significantly more present in patients with PH-LHD. The number of participants with missing values per variable are presented in supplementary table S1.

Validation of the Jacobs score

Using the Jacobs score with a risk score cut-off value of $\geqslant$72 on the whole University Hospitals of Leuven cohort (n=344), PH-LHD was diagnosed in 19%, with a PPV of 100%, a NPV of 55% and 100% specificity. When applied to the internal validation cohort (n=104), it identified 18% (n=9 out of 50) of the patients with PH-LHD without false positives. Specificity was 100%, PPV 100% and NPV 55%.

Logistic regression model

Several predictors were significant (p<0.10) in the univariate logistic regression model (supplementary table S2). The final multivariate logistic regression model included four parameters shown as odds ratio and 95% confidence interval: history of left heart disease, including either coronary artery disease or left valvular heart disease that was worse than mild (OR=12.8, 95% CI 5.2–31.3), left atrial dilation measured by TTE (OR=3.2, 95% CI 1.5–7.1), peak early diastolic (E) flow velocity (mitral E V'_{max}) measured by TTE (OR=1.1, 95% CI 1.0–1.1) and presence of moderate and severe mitral valve regurgitation measured by TTE (OR=16.5, 95% CI 3.8–71.9). The logistic regression model was validated in the independent test set and PH-LHD could be diagnosed in 42% (n=21 out of 50) of the patients, with a PPV of 100%, a NPV of 64% and 100% specificity.
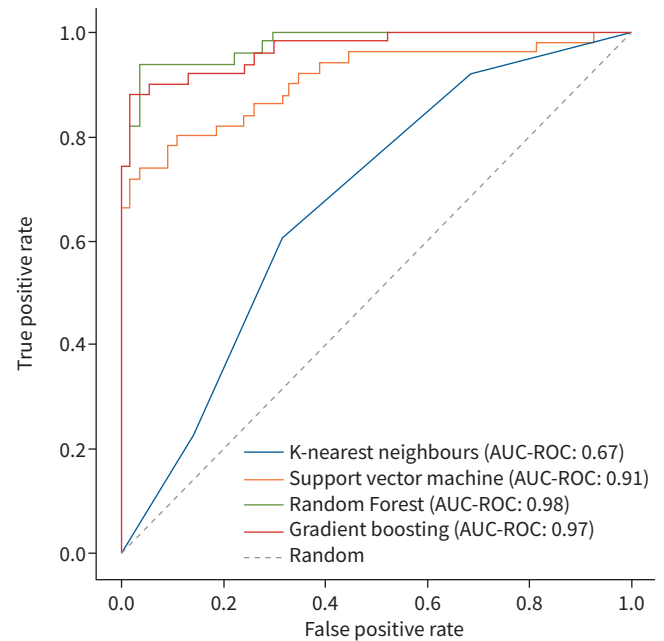
Machine learning model

The best predictive accuracy and AUC-ROC was obtained with the bootstrapped random forest model (figure 1) using 10-fold cross-validation and grid search (parameter grid in supplementary table S3). All potential predictors were used as input for the ML algorithm. Figure 2 shows a plot of the overall accuracy *versus* the number of features, to visualise the variation of accuracy with respect to the number of features. It has been observed that the accuracy quickly increases for the first few features; more specifically, the accuracy quickly increases from 65% to >80% by using the 10 most important parameters only, and stabilises as more features are added. In consultation with the local PH team, the set of input features was reduced to the 20 most important ones. The complete list of these features, ranged according to their Gini impurity-based importance, is shown in supplementary table S4, indicating that higher mitral E V'_{max} measured by TTE, higher mitral valve E/A ratio and the presence of left valvular heart disease worse than mild on echocardiography are the most important predictors for PH-LHD on group level. Other important predictors, listed in the top 15, include the lower haemoglobin level, lower R axis on ECG, lower PRT axes on ECG, higher age, higher left ventricle end diastolic diameter, left atrial dilation worse than mild, lower forced vital capacity and signs of atrial fibrillation on ECG. Additionally, lower forced expiratory volume in 1 s, left valvular disease worse than mild, longer PR interval and higher BMI were included in the top 15 features.

The random forest model, with the reduced set of 20 features, was internally validated in the independent test set. PH-LHD could be correctly diagnosed in 70% (n=35 out of 50) of the patients, with a PPV of 100%, a NPV of 78% and 100% specificity.

To improve the interpretation and the explanatory power of the output from the ML model, the SHAP methodology was used. Figure 3 shows personalised analyses of the model in three exemplative patients. As it performed superior to the logistic regression model, this model was chosen for external validation after being retrained on the entire Leuven cohort.

External validation

Data from 165 subjects were collected from Erasme Hospital Brussels, 91 of whom had PAH and 74 PH-LHD. Their characteristics can be found in table 2. Missing values are reported in supplementary table S1.

FIGURE 1 Comparison of the area under the receiver operating curve (AUC-ROC) from different machine learning techniques. The graph shows the AUC-ROC for a selection of the machine learning techniques on the training cohort of University Hospitals of Leuven (n=240) in order to identify the most discriminative machine learning technique for the clinical question.
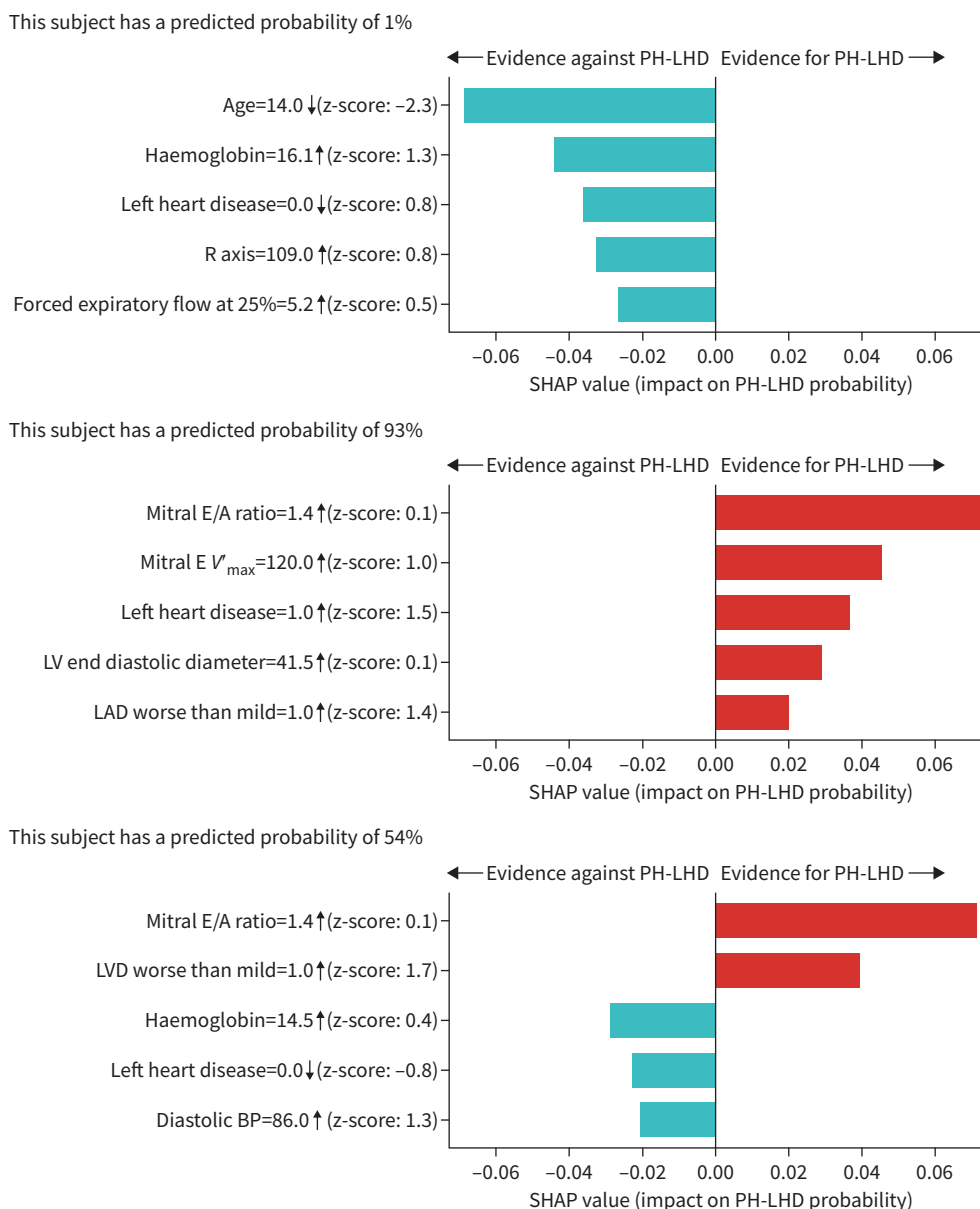
In the external validation cohort, the Jacobs score (cut-off ⩾72) obtained a sensitivity of 3%, PPV of 100%, NPV of 56% and 100% specificity, while the ML model had a sensitivity of 64%, PPV of 100%, NPV of 78% and 100% specificity.

Discussion

To the best of our knowledge, this is the first study investigating noninvasive prediction of PH-LHD with supervised ML, based on the random forest algorithm, and showing that ML significantly improves the sensitivity of PH-LHD prediction at 100% specificity, with the possibility to decrease the number of patients undergoing non-essential invasive diagnostics by 3-fold or more. Out of the 124 included PH-LHD patients (50 in Leuven cohort and 74 in Brussels cohort), 82 were correctly diagnosed (66%) by the ML model meaning that one-third (82 out of 269) of RHCs could have been avoided without therapeutic consequence. With the Jacobs score, 11 out of the 124 PH-LHD patients (9%) were correctly identified.
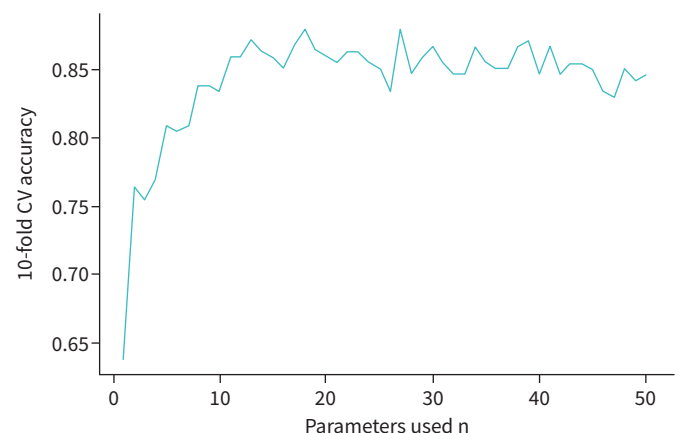
FIGURE 2 Machine learning model performance according to number of features. The plot shows the accuracy of the machine learning model *versus* the number of included features, ordered by increasing predictive value on group level during 10-fold cross-validation on the training cohort of University Hospitals of Leuven (n=240). The accuracy stabilises/mildly decreases when using more than ∼10 features. CV: cross-validation.

This subject has a predicted probability of 1%

This subject has a predicted probability of 93%

This subject has a predicted probability of 54%

FIGURE 3 Individual results of the model in three exemplative patients. Patients 1, 2 and 3 have 1%, 93% and 54% probability, respectively, of having pulmonary hypertension due to left heart disease (PH-LHD), according to the machine learning model. Most contributing factors for this evaluation are displayed on the graph. BP: blood pressure; LAD: left atrial dilation; LV: left ventricular; LVD: left valvular disease; Mitral E/A: mitral valve E/A ratio; Mitral E V'_{max}: the peak early diastolic (E) flow velocity; SHAP value: Shapley Additive Explanations value.

The difficult differential diagnosis between PAH and PH-LHD may be attributed to two aspects [19]. First, various registries showed that 1) the median age of PAH patients at diagnosis is about 65 years or higher and 2) PAH patients are frequently obese and often display comorbidities such as hypertension, diabetes mellitus type II, ischaemic heart disease and atrial fibrillation [20–22], also considered as risk factors for HFpEF. Second, PAWP measurement is fluid dependent, meaning that changes in fluid volume may influence its accurate measurement [23, 24]. Thus, diagnostic decisions exclusively based on PAWP (current gold standard) may be biased.

Over the past years, several prediction scores and a prediction table for PAH or PH-LHD have been developed [1, 2, 4–6, 8, 25]. A common approach to predict PAH or PH-LHD is using logistic regression. However, these models are limited by insufficient predictive value [5, 6], and only two models were prospectively validated [4, 5]. The predictive performance of the Jacobs score in our dataset was similar to

TABLE 2 Characteristics of Erasme Hospital Brussels cohort			
Clinical characteristics	PAH	PH-LHD	p-value
Patients n	91	74	
Age at diagnosis years, median (IQR)	54 (16)	62 (15)	0.003
Male	33 (36)	31 (42)	0.42
BMI $kg·m^{-2}$	31.6±41.7	27.3±7.0	0.38
NYHA class			
I	0 (0)	1 (1)	0.55
II	26 (28)	22 (30)	
III	59 (64)	40 (55)	
IV	7 (8)	10 (14)	
Medical history			
Diabetes mellitus type II	15 (17)	23 (32)	0.04
Hypertension	32 (36)	47 (64)	<0.001
Hypercholesterolaemia	28 (31)	40 (55)	0.002
Obesity	20 (22)	12 (16)	0.43
Valvular surgery without residual valve disease	1 (1)	1 (1)	1.0
Left heart disease	7 (8)	44 (60)	<0.001
Smoking			
Current smoker	14 (16)	5 (7)	0.03
Former smoker	30 (33)	16 (22)	
Never-smoker	46 (51)	51 (71)	
ECG parameters			
Atrial fibrillation	2 (2)	14 (19)	<0.001
SV1+RV6 mm	9.4±4.9	12.0±5.7	0.006
R axis degrees	85.5±65.0	39.9±67.0	<0.001
PR interval ms	169.9±36.3	175.5±37.9	0.40
PRT axes degrees	56.0±20.6	48.8±38.5	0.18
QRS duration ms	93.6±18.7	107.1±31.6	0.002
Echocardiography parameters			
Left atrial dilation worse than mild	11 (12)	13 (18)	0.37
Left valvular disease worse than mild	1 (1)	8 (11)	0.01
Left ventricle end diastolic diameter mm	41.9±7.0	55.7±12.4	<0.001
Mitral valve disease worse than mild	1 (1)	7 (10)	0.005
Mitral E/A ratio	0.9±0.4	1.9±0.9	<0.001
Mitral E V'_{max} $mm·s^{-1}$	57.7±19.9	94.9±27.9	<0.001
Mitral E/E′ ratio	7.9±4.4	15.9±5.6	<0.001
Dilation of the right ventricle			
Not dilated (qualitative)	12 (13)	39 (55)	<0.001
Mildly dilated (qualitative)	40 (44)	25 (35)	
Moderately dilated (qualitative)	8 (9)	0 (0)	
Severely dilated (qualitative)	31 (34)	7 (10)	
Pulmonary function test parameters			
FVC L	2.8±1.0	2.4±0.9	0.01
FEV_1 L	2.2±0.8	1.8±0.7	0.02
FEV_1/FVC ratio	76.8±12.4	78.9±8.1	0.27
Peak expiratory flow $L·s^{-1}$	5.2±1.8	4.5±1.6	0.05
D_{LCO}[#]	NA	NA	
Laboratory blood tests			
Haemoglobin $g·dL^{-1}$	15.0±2.0	12.9±2.0	<0.001
Haematocrit %	45.7±6.1	38.8±6.6	<0.001
Haemodynamic parameters			
Right atrial pressure mmHg	8.6±5.2	11.3±6.2	0.003
Systolic PAP mmHg	82.2±39.2	59.4±16.5	<0.001
Diastolic PAP mmHg	35.8±10.4	29.6±12.4	<0.001
Mean PAP mmHg	50.2±11.2	39.9±10.8	<0.001
CO $L·min^{-1}$	4.0±1.2	4.3±2.3	0.37
CI $L·min^{-1}·m^{-2}$	2.2±0.6	2.2±0.7	0.75
S_{vO_2} %	61.7±11.1	60.3±11.8	0.48
PVR Wood Units	11.4±5.3	4.4±2.6	<0.001

TABLE 2 Continued			
Clinical characteristics	PAH	PH-LHD	p-value
PVR dynes·s·cm^{-5}	910.7±425.4	352.1±206.7	<0.001
PAWP mmHg	9.6±3.5	22.6±6.5	<0.001

Values are expressed as n (%) or mean±SD unless indicated otherwise. PAH: pulmonary arterial hypertension; PH-LHD: pulmonary hypertension due to left heart disease; IQR: interquartile range; BMI: body mass index; NYHA class: New York Heart Association functional class; SV1+RV6: sum of the S deflection in V1 and R deflection in V6 in millimetres on ECG; Mitral E/A: the peak early diastolic (E), atrial systolic (A) transmitral flow velocities for evaluation of left ventricle diastolic function; Mitral E V'_{max}: peak E mitral flow velocity; Mitral E/E′: ratio of early diastolic mitral inflow velocity to early diastolic mitral annulus velocity for the evaluation of left ventricular filling pressure; FVC: forced vital capacity; FEV$_1$: forced expiratory volume in 1 s; FEV$_1$/FVC: Tiffeneau index; D_{LCO}: diffusing capacity of the lung for carbon monoxide; NA: not available; PAP: pulmonary arterial pressure; CO: cardiac output; CI: cardiac index; S_{vO_2}: mixed venous oxygen saturation; PVR: pulmonary vascular resistance; PAWP: pulmonary arterial wedge pressure.

the results of their own prospective validation [3]. The Jacobs score was recently updated in the OPTICS score, *a priori* excluding patients with valve disease, and based on BMI ⩾30, diabetes mellitus type II, atrial fibrillation, dyslipidaemia, history of valvular surgery, sum of SV1 (deflection in V1 in millimetres) and RV6 (deflection in V6 in millimetres) on ECG, and left atrial dilation [26]. The sensitivity for the detection of PH-LHD in our dataset was only 20%, with a specificity of 96% with the OPTICS score. The presently constructed logistic regression model included similar predictive variables as the Jacobs score and had only slightly better predictive performance. This could be expected since our cohort had similar demographic characteristics as the development cohort of the Jacobs score.

When complex relationships between predictors are not captured using conventional regression-based analysis, ML is an appropriate alternative. The random forest model incorporates a larger number of predictors, which are all commonly measured in clinical practice. The model maximised the predictive accuracy and statistical robustness even in a small dataset and can be rapidly performed. The black-box problem, which is commonly seen as an important limitation of the clinical applications of ML, is tackled by using the SHAP methodology. This visualisation method improves understanding of the algorithm by explaining individual predictions by computing the contribution of each feature to the prediction.

Our results are consistent with existing prediction scores, integrating comparable features, including measurements of right and left heart chamber dimensions with estimates of right ventricle and left ventricle filling pressures to discriminate between pre- and post-capillary PH [4, 6, 8]. The emergence of decreased haemoglobin as an important predictor of PH-LHD was more surprising, although iron deficiency is recognised to be common and detrimental in heart failure [27, 28]. Accordingly, BRITTAIN et al. [29] showed higher cell-free haemoglobin in patients with PAH, compared with patients with pulmonary venous hypertension due to ischaemic and non-ischaemic cardiomyopathy with a mPAP >25 mmHg and a PAWP >15 mmHg.

Despite significant differences in patient characteristics between Leuven and Brussels cohorts, including mitral E V'_{max} (first ML predictor; higher in Leuven), the history of left heart disease (third predictor; more common in Leuven) and haemoglobin level (fourth predictor; lower in Leuven), the sensitivity from internal to external validation only dropped by 6 percentage points (from 70% to 64%), showcasing both reproducibility and generalisability for the ML model.

The Jacobs score performed poorly in the external validation set with a sensitivity of 3%, while it did maintain a sensitivity of 19% in the complete University Hospitals of Leuven cohort. A possible explanation for the lower predictive power of the Jacobs score in the Brussels cohort could be the differences in population characteristics. While the patient characteristics of the Jacobs cohort were similar to the characteristics of the Leuven cohort, significant differences were observed when comparing characteristics with the Brussels cohort. The Jacobs score therefore showed good reproducibility but proved to have a lower generalisability to different populations. Despite the limited size of the Leuven cohort (n=344), which was smaller than the cohort used for developing the Jacobs score (n=380), the ML model outperformed the Jacobs score by a significant amount during internal validation (70% *versus* 19% sensitivity) and external validation (64% *versus* 3% sensitivity). Finally, we also applied the H$_2$FPEF score to the Leuven patient cohort. The H$_2$FPEF score is a predictive tool estimating the probability of HFpEF in

patients with unexplained dyspnoea [30, 31], and even though this tool was not designed to distinguish PH-LHD, it has been suggested for this purpose. It includes six predictors: obesity, atrial fibrillation, age >60 years, treatment with two or more antihypertensives, echocardiographic E/E′ ratio >9 and echocardiographic pulmonary artery systolic pressure >35 mmHg. Risk score cut-offs ⩽1, 2–5 and ⩾6 are used to stratify patients into low, intermediate and high HFpEF probability, respectively (range of the score: 0–9 points). Using a cut-off ⩾6/9, the H_2FPEF score identified PH-LHD with a sensitivity of 48%, PPV of 77%, NPV of 62% and specificity of 85%, which is by far inferior to our ML model.

Future investigations are needed to demonstrate efficacy in routine clinical practice. In particular, the potential impact of a web or cloud-based application in secondary care hospitals to reduce the numbers of diagnostic RHC will depend on the pre-test probability of PAH and PH-LHD, which may vary between regions and countries. At this stage we determined only 20 important input features originating from five easily accessible axes (clinical characteristics, laboratory, echocardiography, ECG and spirometry) which are part of the standard clinical workup in such a setting. With the generation of new data, once embedded in clinical practice, the model can be recalibrated to improve upon existing prediction models and suit external populations even better.

Conclusion

The current study demonstrated that a random forest ML model using routinely collected health data is largely superior to existing risk scores and logistic regression models in identifying patients with PH-LHD. With a 100% specific but much more sensitive noninvasive detection tool of PH-LHD, one could avoid numerous invasive RHC in patients with a high pre-test likelihood of PH-LHD, while also improving the positive detection yield for PAH in patients with a predicted likelihood of PAH. The clinical implementation of such tools may decrease burden to patients, change referral strategies to PH centres and reduce economic costs substantially.

Study limitations

Firstly, the used datasets are of limited size, 344 subjects for the Leuven cohort and 165 for the Brussels cohort, given the large number of predictors. However, we have reduced the number of predictors required in order to avoid overfitting, and the results show that even the sizes of the datasets were sufficient to make reliable predictions. A second limitation is that the final diagnosis (PAH *versus* PH-LHD) was based on the clinical gestalt of the local PH team, considering not only the RHC results but also the clinical characteristics, and echocardiographic and laboratory results. This adds a subjective factor to the diagnostic process, increasing the risk of bias especially in patients with overlapping phenotypes. However, we believe that it is unavoidable to include an expert opinion since the interpretation of (borderline) PAWP is complex and influenced by many factors, such as fluid intake, diuresis, respiration and zeroing levels. Only a few patients were classified differently than indicated by PAWP after the review by our PH team (see supplementary table S5). Finally, the ML model was designed in a selected patient population using retrospective data. Patients with PAH and PH-LHD were balanced in the Leuven cohort to counter the real-life imbalance that would heavily bias the prediction model [32]. Instead of oversampling the minority class (synthetic data) or undersampling the majority class (loss of data), we chose to use real-life data to balance the classes. Similar to the University Hospitals of Leuven, Erasme Hospital Brussels is a tertiary centre, which explains the over-representation of patients with PAH, and therefore not reflecting a real-world situation. Before recommending the avoidance of RHC in patients with a low likelihood of PAH, a prospective multicentric study will be conducted in referring centres with a normal PAH prevalence to validate the model.

References

1   Humbert M, Kovacs G, Hoeper MM, *et al.* 2022 ESC/ERS Guidelines for the diagnosis and treatment of pulmonary hypertension: developed by the task force for the diagnosis and treatment of pulmonary hypertension of the European Society of Cardiology (ESC) and the European Respiratory Society (ERS). Endorsed by the International Society for Heart and Lung Transplantation (ISHLT) and the European Reference Network on rare respiratory diseases (ERN-LUNG). *Eur Heart J* 2022; 43: 3618–3731.

2   Vachiéry JL, Tedford RJ, Rosenkranz S, *et al.* Pulmonary hypertension due to left heart disease. *Eur Respir J* 2019; 53: 1801897.

3   Jacobs W, Konings TC, Heymans MW, *et al.* Noninvasive identification of left-sided heart failure in a population suspected of pulmonary arterial hypertension. *Eur Respir J* 2015; 46: 422–430.

4   Bonderman D, Wexberg P, Martischnig AM, *et al.* A noninvasive algorithm to exclude pre-capillary pulmonary hypertension. *Eur Respir J* 2011; 37: 1096–1103.

5   Opotowsky AR, Ojeda J, Rogers F, *et al.* A simple echocardiographic prediction rule for hemodynamics in pulmonary hypertension. *Circ Cardiovasc Imaging* 2012; 5: 765–775.

6   Berthelot E, Montani D, Algalarrondo V, *et al.* A clinical and echocardiographic score to identify pulmonary hypertension due to HFpEF. *J Card Fail* 2017; 23: 29–35.

7   D'Alto M, Romeo E, Argiento P, *et al.* Echocardiographic prediction of pre- *versus* postcapillary pulmonary hypertension. *J Am Soc Echocardiogr* 2015; 28: 108–115.

8   Richter SE, Roberts KE, Preston IR, *et al.* A simple derived prediction score for the identification of an elevated pulmonary artery wedge pressure using precatheterization clinical data in patients referred to a pulmonary hypertension center. *Chest* 2016; 149: 1261–1268.

9   Du Z, Yang Y, Zheng J, *et al.* Accurate prediction of coronary heart disease for patients with hypertension from electronic health records with big data and machine-learning methods: model development and performance evaluation. *JMIR Med Inform* 2020; 8: e17257.

10   Weng SF, Reps J, Kai J, *et al.* Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS One* 2017; 12: e0174944.

11   Taylor RA, Pare JR, Venkatesh AK, *et al.* Prediction of in-hospital mortality in emergency department patients with sepsis: a local big data-driven, machine learning approach. *Acad Emerg Med* 2016; 23: 269–278.

12   Galiè N, Humbert M, Vachiery JL, *et al.* 2015 ESC/ERS Guidelines for the diagnosis and treatment of pulmonary hypertension: The Joint Task Force for the Diagnosis and Treatment of Pulmonary Hypertension of the European Society of Cardiology (ESC) and the European Respiratory Society (ERS). Endorsed by: Association for European Paediatric and Congenital Cardiology (AEPC), International Society for Heart and Lung Transplantation (ISHLT). *Eur Respir J* 2015; 46: 903–975.

13   Rudski LG, Lai WW, Afilalo J, *et al.* Guidelines for the echocardiographic assessment of the right heart in adults: a report from the American Society of Echocardiography endorsed by the European Association of Echocardiography, a registered branch of the European Society of Cardiology, and the Canadian Society of Echocardiography. *J Am Soc Echocardiogr* 2010; 23: 685–713.

14   Copas JB. Regression, prediction and shrinkage. *J R Stat Soc B Methodol* 1983; 45: 311–335.

15   Efron B. Bootstrap methods: another look at the Jackknife. *Ann Stat* 1979; 7: 1–26.

16   Pattengale ND, Alipour M, Bininda-Emonds ORP, *et al.* How many bootstrap replicates are necessary? *In:* Batzoglou S, ed. Research in Computational Molecular Biology (Lecture Notes in Computer Science series). Berlin, Springer, 2009; pp. 184–200.

17   Lundberg S, Lee SI. A unified approach to interpreting model predictions. *arXiv* 2017; preprint [http://arxiv.org/abs/1705.07874].

18   Azur MJ, Stuart EA, Frangakis C, *et al.* Multiple imputation by chained equations: what is it and how does it work?: multiple imputation by chained equations. *Int J Methods Psychiatr Res* 2011; 20: 40–49.

19   Opitz CF, Hoeper MM, Gibbs JSR, *et al.* Pre-capillary, combined, and post-capillary pulmonary hypertension. *J Am Coll Cardiol* 2016; 68: 368–378.

20 Ling Y, Johnson MK, Kiely DG, *et al.* Changing demographics, epidemiology, and survival of incident pulmonary arterial hypertension: results from the pulmonary hypertension registry of the United Kingdom and Ireland. *Am J Respir Crit Care Med* 2012; 186: 790–796.

21 Hoeper MM, Huscher D, Ghofrani HA, *et al.* Elderly patients diagnosed with idiopathic pulmonary arterial hypertension: results from the COMPERA registry. *Int J Cardiol* 2013; 168: 871–880.

22 Rådegran G, Kjellström B, Ekmehag B, *et al.* Characteristics and survival of adult Swedish PAH and CTEPH patients 2000-2014. *Scand Cardiovasc J* 2016; 50: 243–250.

23 D'Alto M, Romeo E, Argiento P, *et al.* Clinical relevance of fluid challenge in patients evaluated for pulmonary hypertension. *Chest* 2017; 151: 119–126.

24 Borlaug BA, Paulus WJ. Heart failure with preserved ejection fraction: pathophysiology, diagnosis, and treatment. *Eur Heart J* 2011; 32: 670–679.

25 Dardi F, Tanese N, Caravita S, *et al.* P4542 A new score to differentiate idiopathic pulmonary arterial hypertension from pulmonary hypertension due to heart failure with preserved ejection fraction. *Eur Heart J* 2018; 39: Suppl_1, ehy563.P4542.

26 Jansen SMA, Huis in 't Veld Anna E, Jacobs W, *et al.* Noninvasive prediction of elevated wedge pressure in pulmonary hypertension patients without clear signs of left-sided heart disease: external validation of the OPTICS risk score. *J Am Heart Assoc* 2020; 9: e015992.

27 Rhodes CJ, Wharton J, Howard L, *et al.* Iron deficiency in pulmonary arterial hypertension: a potential therapeutic target. *Eur Respir J* 2011; 38: 1453–1460.

28 Humbert M, McLaughlin V, Gibbs JSR, *et al.* Sotatercept for the treatment of pulmonary arterial hypertension. *N Engl J Med* 2021; 384: 1204–1215.

29 Brittain EL, Janz DR, Austin ED, *et al.* Elevation of plasma cell-free hemoglobin in pulmonary arterial hypertension. *Chest* 2014; 146: 1478–1485.

30 Reddy YNV, Carter RE, Obokata M, *et al.* A simple, evidence-based approach to help guide diagnosis of heart failure with preserved ejection fraction. *Circulation* 2018; 138: 861–870.

31 Kianzad A, van Wezenbeek J, Celant LR, *et al.* Idiopathic pulmonary arterial hypertension patients with a high H2FPEF-score: insights from the Amsterdam UMC PAH-cohort. *J Heart Lung Transplant* 2022; 41: 1075–1085.

32 Fernández A, García S, Galar M, *et al.* Learning from Imbalanced Data Sets. Cham, Springer International Publishing, 2018.