ORIGINAL ARTICLE
PULMONARY VASCULAR DISEASE

Identifying early pulmonary arterial hypertension biomarkers in systemic sclerosis: machine learning on proteomics from the DETECT cohort

Yasmina Bauer[1,2], Simon de Bernard[3], Peter Hickey[4,5], Karri Ballard[6], Jeremy Cruz[6], Peter Cornelisse[2], Harbajan Chadha-Boreham[7], Oliver Distler[8], Daniel Rosenberg[7], Martin Doelberg[7], Sebastien Roux[2], Oliver Nayler[2] and Allan Lawrie[4]

Affiliations: [1]Galapagos GmbH, Basel, Switzerland. [2]Idorsia Pharmaceuticals Ltd, Allschwil, Switzerland. [3]AltraBio, Lyon, France. [4]Dept of Infection, Immunity and Cardiovascular Disease, Medical School, University of Sheffield, Sheffield, UK. [5]Sheffield Pulmonary Vascular Disease Unit, Royal Hallamshire Hospital, Sheffield, UK. [6]Myriad RBM, Austin, TX, USA. [7]Actelion Pharmaceuticals Ltd, Allschwil, Switzerland. [8]Dept of Rheumatology, University Hospital Zurich, Zurich, Switzerland.

Correspondence: Allan Lawrie, Dept of Infection, Immunity and Cardiovascular Disease, Medical School, University of Sheffield, Beech Hill Road, Sheffield, S10 2RX, UK. E-mail: a.lawrie@sheffield.ac.uk

@ERSpublications
Early screening for pulmonary arterial hypertension in patients with systemic sclerosis improves patient outcome. This study identified a novel eight-protein biomarker panel that has the potential to assist early detection of PAH in this patient group. https://bit.ly/373BNkL

Cite this article as: Bauer Y, de Bernard S, Hickey P, *et al.* Identifying early pulmonary arterial hypertension biomarkers in systemic sclerosis: machine learning on proteomics from the DETECT cohort. *Eur Respir J* 2021; 57: 2002591 [https://doi.org/10.1183/13993003.02591-2020].

ABSTRACT Pulmonary arterial hypertension (PAH) is a devastating complication of systemic sclerosis (SSc). Screening for PAH in SSc has increased detection, allowed early treatment for PAH and improved patient outcomes. Blood-based biomarkers that reliably identify SSc patients at risk of PAH, or with early disease, would significantly improve screening, potentially leading to improved survival, and provide novel mechanistic insights into early disease. The main objective of this study was to identify a proteomic biomarker signature that could discriminate SSc patients with and without PAH using a machine learning approach and to validate the findings in an external cohort.

Serum samples from patients with SSc and PAH (n=77) and SSc without pulmonary hypertension (non-PH) (n=80) were randomly selected from the clinical DETECT study and underwent proteomic screening using the Myriad RBM Discovery platform consisting of 313 proteins. Samples from an independent validation SSc cohort (PAH n=22 and non-PH n=22) were obtained from the University of Sheffield (Sheffield, UK).

Random forest analysis identified a novel panel of eight proteins, comprising collagen IV, endostatin, insulin-like growth factor binding protein (IGFBP)-2, IGFBP-7, matrix metallopeptidase-2, neuropilin-1, N-terminal pro-brain natriuretic peptide and RAGE (receptor for advanced glycation end products), that discriminated PAH from non-PH in SSc patients in the DETECT Discovery Cohort (average area under the receiver operating characteristic curve 0.741, 65.1% sensitivity/69.0% specificity), which was reproduced in the Sheffield Confirmatory Cohort (81.1% accuracy, 77.3% sensitivity/86.5% specificity).

This novel eight-protein biomarker panel has the potential to improve early detection of PAH in SSc patients and may provide novel insights into the pathogenesis of PAH in the context of SSc.

This article has an editorial commentary: https://doi.org/10.1183/13993003.00205-2021

This article has supplementary material available from erj.ersjournals.com

Received: 6 July 2020 | Accepted: 17 Nov 2020

Introduction

Pulmonary arterial hypertension (PAH) is a devastating complication of systemic sclerosis (SSc) affecting 7–12% of patients with this condition [1, 2]. Right heart failure as a result of PAH is one of the leading causes of death in this patient cohort, accounting for 26% of deaths [3], and SSc-PAH represents 15–20% of all forms of PAH in Europe [4] with a 30% 1-year mortality [5, 6]. There is significant interest in developing improved screening tools using a variety of approaches including blood biomarkers, imaging, exercise testing (reviewed in [7]) and real-world healthcare resource utilisation data [8, 9] to improve the detection of PAH and decrease the time from first symptom to diagnosis [10]. Study data from Humbert *et al.* [11] demonstrate that screening for PAH in an at-risk population of patients with SSc enables early diagnosis and therefore treatment of a milder form of the disease. This resulted in a significant increase in survival, but it is important to acknowledge the lead-time and length-time biases introduced by the screening study approach [11], and earlier treatment remains an important goal for PAH. To this end, the development of the DETECT algorithm, which has been shown to outperform symptom- and transthoracic echocardiography-based diagnosis in a subset of SSc patients with early stages of PAH [12, 13], has been proposed in the diagnosis work-up of the European Society of Cardiology/European Respiratory Society guidelines [14].

The DETECT algorithm includes values for the circulating biomarkers N-terminal pro-brain natriuretic peptide (NT-proBNP) and uric acid, demonstrating the potential of biomarkers to support early detection of PAH in SSc patients. Utilising serum samples and clinical data collected during the DETECT study, we hypothesised that a broader proteomic signature could be developed to classify patients with SSc into those with and without PAH. Using serum samples from 157 randomly selected patients from the DETECT study (DETECT Discovery Cohort) we identified an eight-protein signature for PAH in SSc patients. This panel was subsequently reproduced in an independent cohort of 44 sequentially consented SSc patients (Sheffield Confirmatory Cohort). We also demonstrate that some protein biomarkers can predict individual clinical variables from the DETECT study.

Materials and methods

DETECT Discovery Cohort

Patients were eligible for inclusion in DETECT if they were aged ⩾18 years and had 1) a diagnosis of SSc of >3 years duration from the first non-Raynaud's symptom, 2) diffusing capacity of the lung for carbon monoxide (D_{LCO}) <60% predicted, 3) forced vital capacity (FVC) ⩾40% predicted and 4) not had pulmonary hypertension (PH) confirmed by right heart catheterisation (RHC) prior to enrolment. PAH was subsequently confirmed by RHC during the DETECT study. Approval was obtained from all relevant boards/ethics committees and all patients provided written informed consent [12]. Blood sampling was performed within a few days of RHC. All samples were collected and processed following Standard Operating Procedures and stored at −80°C until assaying. Stored serum samples from 77 out of 87 patients from the DETECT study with World Health Organization Group 1 PH (PAH) [3] were used in the DETECT Discovery Cohort. 10 patients with PAH were excluded because the samples reached the time limit of storage by consent. A random sample of 80 out of 321 non-PH DETECT study patients was selected for controls.

Sheffield Confirmatory Cohort

An independent validation cohort was obtained from PAH treatment-naive patients sequentially recruited to The Sheffield Teaching Hospitals Observational Study of Patients with Pulmonary Hypertension, Cardiovascular and Lung Disease (STH-ObS) undergoing RHC for suspected PH at the Sheffield Pulmonary Vascular Disease Unit (Royal Hallamshire Hospital, Sheffield, UK) (Research Ethics Committee 18/YH/0441). In addition to RHC, patients were systematically investigated with high-resolution computed tomography, cardiac magnetic resonance imaging, pulmonary function testing and the incremental shuttle walk test according to local standard procedure. Serum samples were collected from diagnostic RHC prior to diagnosis of PH from 2008 to 2015 following local Standard Operating Procedures. All patients had a clear diagnosis of SSc and investigations consistent with PAH without any concomitant interstitial lung disease (PAH n=22). A cohort of disease controls was selected from patients who had a firm diagnosis of SSc, but in whom RHC excluded PH (mean pulmonary arterial pressure (mPAP) <25 mmHg) (non-PH n=22). Serum from all patients was aliquoted and stored at −80°C until requested for this study.

Measurement of circulating protein biomarkers

All serum samples were stored at less than −70°C until tested. Samples were thawed at room temperature, vortexed, spun at 3700×*g* for 5 min to remove precipitates and transferred to a master microtitre plate. 313 analytes were then assessed in serum using a multiplexed immunoassay (DiscoveryMAP version 3.0 assay;

Myriad RBM, Austin, TX, USA). Monitoring of internal controls and batch testing between cohorts was performed internally by Myriad RBM.

Clinical variables

The demographic and clinical variables selected from the DETECT study [3] to explore their relationship with the serum biomarkers included 1) demographic and clinical characteristics: sex, age, body mass index, smoking history, SSc subtype (limited, diffuse and mixed), current/past telangiectasias, disease severity (modified Rodnan skin score (mRSS)); 2) echocardiography: qualitative assessment of right ventricle pump function (normal, slightly impaired, moderately impaired and severely impaired), right atrium area, right ventricle area, right ventricle diameter, tricuspid annular plane systolic excursion (TAPSE) and tricuspid regurgitant velocity; 3) electrocardiography: right axis deviation; 4) RHC: pulmonary vascular resistance (PVR), mPAP and pulmonary arterial wedge pressure (PAWP); 5) pulmonary function tests: FVC, FVC % pred, D_{LCO} and D_{LCO} % pred; and 6) NT-proBNP.

These variables were also selected from the Sheffield Confirmatory Cohort, with the exception of current/past telangiectasias, disease severity (mRSS), right ventricle area, right ventricle diameter and TAPSE.

Data processing and analysis

The selection of patients for the DETECT Discovery Cohort was performed using all samples from patients with PAH who had valid serum samples followed by exclusion of patients for pragmatic reasons (quality control) and a random selection from the non-PH group. The Sheffield Confirmatory Cohort comprised patients sequentially recruited with suspected PH during the period from 2008 to 2015 who underwent extensive phenotyping for suspected PAH. A subselection for patients who had SSc either with or without PAH was then performed.

The demographic and clinical variables for the DETECT Discovery Cohort (table 1) and the Sheffield Confirmatory Cohort (table 2) were summarised using descriptive statistics, and the PAH and non-PH groups were compared using the Wilcoxon rank-sum test for continuous variables (and for the "qualitative evaluation of right ventricle pump" after conversion of the ordered factor levels to 1–4 values) and Fisher's exact test for categorical variables.

Analytes with >50% missing values or near zero variance were excluded from analyses. One patient from the Sheffield Confirmatory Cohort was removed because the data contained 25 missing values for analytes. Out of the 313 analytes that were measured, 271 analytes were used for the DETECT Discovery Cohort (n=157) and 258 analytes were used for the Sheffield Confirmatory Cohort (n=44) for subsequent analysis. Data transformations were applied to some of the clinical variables in order to obtain normal distributions. A $\log(x)$ transformation was applied to the variable rhcpvr (PVR), and a $\log(1+x)$ transformation was applied to the variables mrsstot (mRSS) and mmraa (right atrium area). After initial analyses we found that age was the 12th most important variable in the DETECT Discovery Cohort and 26th in the Sheffield Confirmatory Cohort. Sex was not in the top 100 variables for either cohort. We therefore chose not to correct for age or sex in subsequent analyses.

Random forests

Missing values were imputed with the NIPALS algorithm (www.github.com/kwstat/nipals; R package version 0.5). Within the random forest (RF) analysis, the number of variables randomly sampled as candidates at each node was set to its default value. The importance of biomarker variables in each cohort was assessed using the mean decrease in the node impurity criterion (Gini index). Three distinct RF analyses were performed with the aim of identifying a consistent panel of biomarkers to classify PAH from non-PH in patients with SSc: 1) in the DETECT Discovery Cohort using the 271 analytes detected, 2) in the Sheffield Confirmatory Cohort using the 258 analytes detected, and 3) in both cohorts using 238 common protein analytes between the discovery and validation datasets (31 analytes from the DETECT dataset could not be mapped to the Sheffield dataset and 56 analytes from the Sheffield dataset could not be mapped to the DETECT dataset, because they were either previously filtered or absent in the original dataset). Further RF analyses performed all possible combinations of smaller panels of the eight or less biomarkers that were common to the top 20 biomarkers in the three RF analyses. Performance of the RFs was internally validated by averaging the area under the receiver operating characteristic curve (ROC-AUC) analyses of repeated (100 times) 10-fold cross-validations. RF analyses were performed using the package randomForest version 4.6-14 and R version 3.5.0 (R Project, Vienna, Austria) from CRAN (www.CRAN.R-project.org). The performance of the RF was assessed in the Sheffield Confirmatory Cohort using balance accuracy.

TABLE 1 Baseline characteristics of pulmonary arterial hypertension (PAH) and non-pulmonary hypertension (non-PH) systemic sclerosis (SSc) patients in the DETECT Discovery Cohort

	Non-PH	PAH	All
All patients	80	77	157
Demographic and clinical parameters			
Age years			
Patients	80	76	156
Mean±SD	55.7±11.5	61.5±9.5	58.5±11.0
Median (IQR)	57.0 (48.0–61.0)	62.0 (55.8–68.0)	59.0 (51.0–66.0)
Minimum–maximum	26–82	39–80	26–82
		p=0.00059	
Sex			
Patients	80	77	157
Female	71 (88.8)	58 (75.3)	129 (82.2)
Male	9 (11.2)	19 (24.7)	28 (17.8)
		p=0.037	
BMI kg·m^{-2}			
Patients	79	76	155
Mean±SD	26.4±7.0	26.3±5.7	26.3±6.4
Median (IQR)	24.6 (21.9–29.1)	25.6 (22.4–28.7)	25.1 (22.1–28.8)
		p=0.73	
Smoking history	35/80 (43.8)	38/77 (49.4)	73/157 (46.5)
		p=0.52	
SSc subtype			
Patients	80	76	156
Limited	46 (57.5)	55 (72.4)	101 (64.7)
Diffuse	28 (35.0)	15 (19.7)	43 (27.6)
Mixed	6 (7.5)	6 (7.9)	12 (7.7)
		p=0.1	
Current/past telangiectasias	59/80 (73.8)	67/77 (87.0)	126/157 (80.3)
		p=0.045	
Disease severity mRSS			
Patients	80	76	156
Mean±SD	9.4±8.2	11.2±8.2	10.3±8.3
Median (IQR)	6.5 (4.0–11.2)	9.0 (5.8–14.2)	8.0 (4.8–13.2)
		p=0.047	
Echocardiography			
Qualitative evaluation of right ventricle pump			
Patients	80	77	157
Normal	79 (98.8)	60 (77.9)	139 (88.5)
Slightly impaired	1 (1.2)	8 (10.4)	9 (5.7)
Moderately impaired	0 (0)	5 (6.5)	5 (3.2)
Severely impaired	0 (0)	4 (5.2)	4 (2.5)
		p=4.2e-05	
Right atrium area (maximum) cm^2			
Patients	75	73	148
Mean±SD	13.4±4.6	16.8±5.9	15.1±5.5
Median (IQR)	13.0 (10.2–15.3)	15.5 (13.9–19.0)	14.0 (11.9–16.9)
		p=2.7e-05	
Right ventricle area (maximum) cm^2			
Patients	76	73	149
Mean±SD	15.1±5.2	19.4±6.8	17.2±6.4
Median (IQR)	14.2 (12.0–17.1)	18.6 (13.6–22.7)	16.0 (13.0–20.8)
		p=5.8e-05	
Right ventricle diameter (maximum) mm			
Patients	77	73	150
Mean±SD	29.0±8.6	32.6±9.0	30.8±8.9
Median (IQR)	29.0 (25.0–32.0)	32.0 (27.0–36.3)	30.0 (26.0–35.0)
		p=0.002	

Continued

TABLE 1 Continued

	Non-PH	PAH	All
TAPSE m			
Patients	74	67	141
Mean±SD	23.1±4.4	21.2±4	22.2±4.3
Median (IQR)	23.2 (20.0–26.0)	21.0 (19.0–24.0) p=0.0076	22.0 (19.0–25.0)
TRV $m \cdot s^{-1}$			
Patients	75	74	149
Mean±SD	2.4±0.5	3.0±0.7	2.7±0.7
Median (IQR)	2.4 (2.1–2.7)	2.9 (2.5–3.5) p=9.8e-09	2.6 (2.3–3.1)
Electrocardiography			
Presence of right axis deviation	2/74 (2.7)	10/75 (13.3) p=0.031	12/149 (8.1)
Right heart catheterisation			
PVR $dyn \cdot s \cdot cm^{-5}$			
Patients	78	77	155
Mean±SD	141.1±60.9	366.0±207.1	252.8±189.0
Median (IQR)	144.6 (100.0–177.3)	300.0 (226.9–410.3) p=2.8e-20	208.2 (136.1–296.9)
mPAP mmHg			
Patients	80	77	157
Mean±SD	17.9±3.7	32.3±7.8	25.0±9.4
Median (IQR)	18.0 (15.8–21.0)	29.0 (27.0–36.0) p=2.7e-27	24.0 (18.0–29.0)
PAWP mmHg			
Patients	79	77	156
Mean±SD	9.1±3.4	10.1±3.2	9.6±3.3
Median (IQR)	9.0 (7.0–11.5)	11.0 (8.0–12.0) p=0.034	10.0 (7.0–12.0)
$\log_{10}$(NT-proBNP) $pg \cdot mL^{-1}$			
Patients	80	77	157
Mean±SD	2.5±0.6	2.9±0.5	2.7±0.6
Median (IQR)	2.6 (2.2–2.9)	3.0 (2.6–3.2) p=4.5e-05	2.7 (2.3–3.0)
Pulmonary function tests			
FVC % pred			
Patients	80	77	157
Mean±SD	82.2±20.0	90.9±18.2	86.4±19.6
Median (IQR)	82.0 (65.7–96.5)	89.0 (77.4–101.0) p=0.0075	84.2 (72.0–99.1)
D_{LCO} % pred			
Patients	80	77	157
Mean±SD	47.4±8.7	43.2±10.7	45.4±9.9
Median (IQR)	49.1 (41.1–54.1)	44.0 (36.0–53.0) p=0.016	47.3 (39.0–53.2)

Data are presented as n, n (%) or n/N (%), unless otherwise stated. IQR: interquartile range; BMI: body mass index; mRSS: modified Rodnan skin score; TAPSE: tricuspid annular plane systolic excursion; TRV: tricuspid regurgitant velocity; PVR: pulmonary vascular resistance; mPAP: mean pulmonary arterial pressure; PAWP: pulmonary arterial wedge pressure; NT-proBNP: N-terminal pro-brain natriuretic peptide; FVC: forced vital capacity; D_{LCO}: diffusing capacity of the lung for carbon monoxide. Comparison of PAH and non-PH groups performed using the Wilcoxon rank-sum test for continuous variables (and for the "qualitative evaluation of right ventricle pump" after conversion of the ordered factor levels to 1–4 values) and Fisher's exact test for categorical variables.

Partial least squares

17 clinical variables from the DETECT Discovery Cohort were taken individually and assessed as a dependent variable and sparse partial least squares (SPLS) regressions performed to identify biomarkers best able to predict the clinical variable. The biomarker data were Box–Cox transformed to normalise the residual distributions. Lambda parameters ranging from −2 to 2 were allowed. Performances were estimated by repeated (100 times) 10-fold cross-validations. The number of components was set to 1 and

TABLE 2 Patient characteristics of pulmonary arterial hypertension (PAH) and non-pulmonary hypertension (non-PH) systemic sclerosis (SSc) patients in the Sheffield Confirmatory Cohort

	Non-PH	PAH	All
All patients	22	22	44
Demographic and clinical parameters			
Age years			
Patients	22	22	44
Mean±SD	62.3±9.7	68.1±8.0	65.2±9.2
Median (IQR)	61.5 (56.3–67.8)	69.0 (63.3–72.0)	64.5 (58.8–72.0)
Minimum–maximum	45–84	53–84	45–84
		p=0.026	
Sex			
Patients	22	22	44
Female	20 (90.9)	15 (68.2)	35 (79.5)
Male	2 (9.1)	7 (31.8)	9 (20.5)
		p=0.091	
BMI kg·m^{-2}			
Patients	22	20	42
Mean±SD	27.7±5.8	27.5±6.9	27.6±6.3
Median (IQR)	28.1 (23.9–32.6)	28.1 (21.7–31.7)	28.1 (21.9–32.0)
		p=0.782	
Smoking history	8/19 (42.1)	13/19 (68.4)	21/38 (55.3)
		p=1	
SSc subtype			
Patients	22	22	44
Limited	21 (95.5)	22 (100)	43 (97.7)
Diffuse	1 (4.5)	0 (0)	1 (2.3)
Mixed	0 (0)	0 (0)	0 (0)
		p=1	
Echocardiography/CTPA/cardiac MRI			
Qualitative evaluation of right ventricle pump			
Patients	22	22	44
Normal	18 (81.8)	11 (50.0)	29 (65.9)
Mild impairment	4 (18.2)	2 (9.1)	6 (13.6)
Moderate impairment	0 (0)	1 (4.5)	1 (2.3)
Severe impairment	0 (0)	8 (36.4)	8 (18.2)
		p=0.0083	
Right atrium area (maximum) cm^2			
Patients	22	22	44
Mean±SD	16.0±5.4	22.5±9.6	19.2±8.4
Median (IQR)	15.4 (11.5–17.1)	19.8 (13.7–29.9)	16.9 (13.2–23.7)
		p=0.018	
TRV m·s^{-1}			
Patients	17	22	39
Mean±SD	2.7±0.3	3.7±0.8	3.2±0.8
Median (IQR)	2.7 (2.5–2.8)	3.4 (3.1–4.2)	3.1 (2.7–3.5)
		p=1.98e-05	
Electrocardiography			
Presence of right axis deviation	1/20 (5.0)	9/22 (40.9)	10/42 (23.8)
		p=0.45	
Right heart catheterisation			
PVR dyn·s·cm^{-5}			
Patients	22	22	44
Mean±SD	146.4±56.5	548.2±375.4	347.3±334.2
Median (IQR)	139.5 (116.8–170.8)	396.0 (271.2–850.2)	219.0 (142.8–395.5)
		p=1.5e-09	
mPAP mmHg			
Patients	22	22	44
Mean±SD	20.7±3.1	41.7±15.3	31.2±15.2
Median (IQR)	21.0 (19.0–22.0)	38.0 (30.0–52.0)	26.0 (21.0–37.0)
		p=3.0e-08	

Continued

TABLE 2 Continued

	Non-PH	PAH	All
PAWP mmHg			
Patients	22	22	44
Mean±SD	9.1±3.8	11.7±4.9	10.4±4.5
Median (IQR)	8.5 (7.0–11.8)	12.0 (8.0–14.0) p=0.071	10.0 (7.8–13.2)
log_{10}(NT-proBNP) pg·mL^{-1}			
Patients	22	22	44
Mean±SD	2.6±0.5	3.3±0.5	3.0±0.6
Median (IQR)	2.7 (2.4–2.9)	3.3 (3.0–3.7) p=9.3e-05	2.9 (2.6–3.3)
Pulmonary function tests			
FVC % pred			
Patients	22	22	44
Mean±SD	90.8±22.6	101.2±15.7	96.0±20.0
Median (IQR)	94.7 (73.4–107.9)	100.8 (90.9–111.5) p=0.13	96.3 (80.9–110.9)
D_{LCO} % pred			
Patients	21	22	43
Mean±SD	56.2±13.4	43.9±12.7	49.9±14.3
Median (IQR)	55.1 (46.7–67.3)	43.6 (39.6–47.0) p=0.002	46.7 (42.0–56.5)

Data are presented as n, n (%) or n/N (%), unless otherwise stated. IQR: interquartile range; BMI: body mass index; CTPA: computed tomography pulmonary angiography; MRI: magnetic resonance imaging; TRV: tricuspid regurgitant velocity; PVR: pulmonary vascular resistance; mPAP: mean pulmonary arterial pressure; PAWP: pulmonary arterial wedge pressure; NT-proBNP: N-terminal pro-brain natriuretic peptide; FVC: forced vital capacity; D_{LCO}: diffusing capacity of the lung for carbon monoxide. Comparison of PAH and non-PH groups performed using the Wilcoxon rank-sum test for continuous variables (and for the "qualitative evaluation of right ventricle pump" after conversion of the ordered factor levels to 1–4 values) and Fisher's exact test for categorical variables.

the sparsity level at 0.7. Missing values were imputed within the cross-validation process using a k-nearest-neighbours approach (k=5). SPLS analyses were performed using the package spls version 2.2-3 run in R version 3.5.0 from CRAN.

Results

Patient disposition, demographic and clinical characteristics

The patient disposition flowchart (figure 1) shows the number of patients and analytes for the DETECT Discovery Cohort and the Sheffield Confirmatory Cohort. The demographic and clinical characteristics of the two cohorts are summarised and the two groups compared as shown in tables 1 and 2, respectively. A comparison of the two cohorts is included as supplementary figures S1–S9. This comparison confirmed, as expected, that the DETECT samples (collected from rheumatology clinics) were slightly more heterogeneous with less advanced PAH than the Sheffield samples collected from a specialist PH referral centre.

Protein biomarker selection using RFs

Analyses of serum samples from PAH (n=77) and non-PH patients (n=80) from the DETECT Discovery Cohort identified 271 of the 313 protein analytes on the Myriad RBM Discovery platform that passed quality control (supplementary table S1). RF analysis identified proteins that segregated PAH from non-PH patients with SSc with an average ROC-AUC of 0.71. Figure 2a shows the top 20 variables (proteins) of importance to distinguish PAH, as ranked by the mean decrease in the Gini index. To determine whether this could be replicated in a distinct cohort of treatment-naive patients with samples collected at diagnostic RHC, we ran 44 serum samples (PAH n=22 and no-PH n=22) from Sheffield on the same Myriad RBM Discovery platform. In this cohort, 258 out of 313 protein analytes passed quality control (supplementary table S1). An independent RF analysis identified proteins that predicted PAH with a ROC-AUC of 0.83. Figure 2b shows the top 20 variables (proteins) of importance in the Sheffield Confirmatory Cohort.

Encouragingly, 238 common analytes were consistently measured in both the DETECT and Sheffield cohorts, and an accuracy of 86% was observed when applying the RF trained on the DETECT Discovery

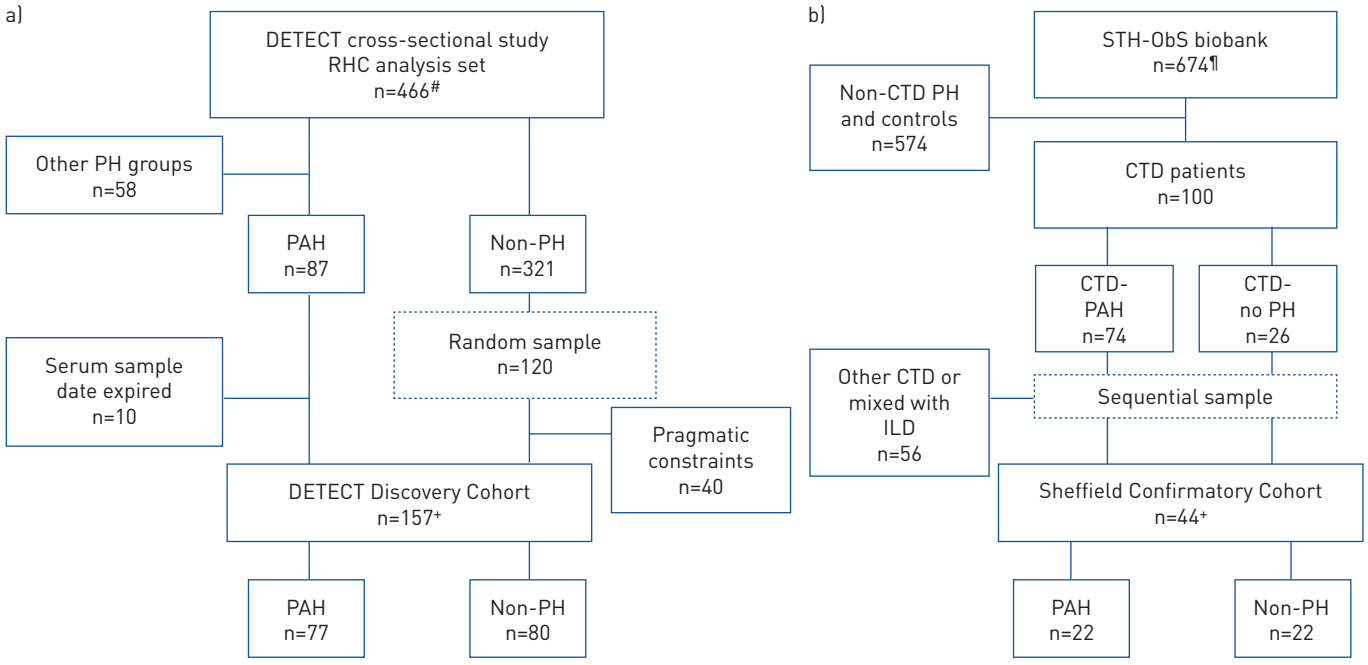

FIGURE 1 Patients and analytes for a) the DETECT Discovery Cohort and b) the Sheffield Confirmatory Cohort. RHC: right heart catheterisation; PH: pulmonary hypertension; PAH: pulmonary arterial hypertension; STH-ObS: The Sheffield Teaching Hospitals Observational Study of Patients with Pulmonary Hypertension, Cardiovascular and Lung Disease; CTD: connective tissue disease; ILD: interstitial lung disease. #: enrolled 2008–2011; ¶: enrolled 2008–2015; +: 271 protein analytes passed quality control in the DETECT Discovery Cohort and 238 protein analytes passed quality control in the Sheffield Confirmatory Cohort (238 protein analytes were suitable for investigation in both cohorts).

Cohort to the Sheffield Confirmatory Cohort. Specifically, collagen IV, endostatin, insulin-like growth factor binding protein (IGFBP)-2, IGFBP-7, matrix metallopeptidase (MMP)-2, neuropilin-1, N-terminal pro-brain natriuretic peptide (NT-proBNP) and RAGE (receptor for advanced glycation end products) were identified as common PAH biomarkers. All results and summary statistics of the biomarkers that were analysed in the DETECT Discovery Cohort and in the Sheffield Confirmatory Cohort are shown in supplementary tables S2 and S3, respectively. Although eight of the top 20 variables of importance were identified in both cohorts after independent RF analysis, a different ranking of the common biomarkers in the two cohorts was noted (figure 2a and b). We therefore performed a new RF analysis on the DETECT Discovery Cohort dataset using the 238 common analytes identified in both cohorts. The 20 top-most important variables of this RF analysis are shown in figure 2c. The eight common PAH biomarkers from

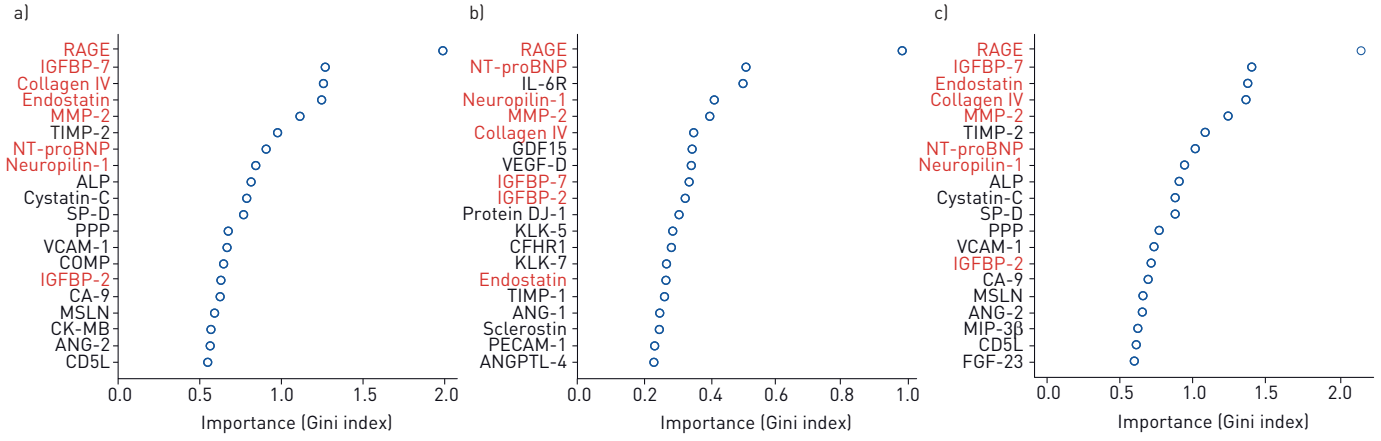

FIGURE 2 Variables (proteins) of importance to classify pulmonary arterial hypertension. Variable importance output of random forests applied to a) the DETECT Discovery Cohort, b) the Sheffield Confirmatory Cohort and c) 238 common proteins between the two cohorts, applied on the DETECT Discovery Cohort. The plots show the most important variables (*y*-axis) as assessed by the mean decrease of the Gini index (*x*-axis). Proteins are ordered top to bottom as most to least important. The eight common variables in all analyses appear in red. See supplementary table S1 for details of the proteins on the Myriad RBM Discovery platform.

the independent analysis of each cohort were again selected in the top 20 variables of importance. The average ROC-AUC for this new analysis in the DETECT Discovery Cohort was 0.72.

The individual protein levels of the conserved eight biomarkers were significantly higher in SSc patients with PAH compared with non-PH patients in the DETECT Discovery Cohort and the Sheffield Confirmatory Cohort (figure 3).

Performance of the eight-protein panel to classify PAH

To determine the potential of the eight protein biomarkers to classify PAH from a mixed cohort of patients with SSc, we performed further RF analyses for all 255 possible combinations of the eight biomarkers identified to determine the panel with the best performance. Panel performance was estimated by repeated cross-validation and a subset panel of six biomarkers, including RAGE, IGFBP-7, collagen IV, endostatin, MMP-2 and IGFBP-2, classified PAH with the best ROC-AUC (0.751) in the DETECT Discovery Cohort with a sensitivity of 66.8% and a specificity of 71.4% (figure 4a). We next assessed the performance of this six-protein biomarker panel in the Sheffield Confirmatory Cohort, which gave a ROC-AUC of 0.866 (figure 4b) and balanced accuracy of 0.705 with a sensitivity of 54.5% and a specificity of 86.4%.

Given the decrease in sensitivity observed with our six-protein biomarker panel in the Sheffield Confirmatory Cohort, we tested whether adding back NT-proBNP or NT-proBNP plus neuropilin-1 (since NT-proBNP is already part of the DETECT algorithm) would improve the reproducibility of the panel. As expected from our previous analysis, adding NT-proBNP (seven-biomarker panel; figure 4c) or NT-proBNP plus neuropilin-1 (eight-biomarker panel; figure 4d) produced a reduced performance in the DETECT Discovery Cohort, generating a ROC-AUC of 0.741 with a sensitivity of 65.2% and a specificity of 68.9% for the seven-biomarker panel (figure 4c) and a ROC-AUC of 0.741 with a sensitivity of 65.1% and a specificity of 69.0% for the eight-biomarker panel (figure 4d). We next tested the performance of both the seven- and eight-protein panels in the Sheffield Confirmatory Cohort. For the seven-protein panel including NT-proBNP we achieved a balanced accuracy of 0.77 with a sensitivity of 68.2% and a specificity of 86.4%. This was slightly improved in the eight-protein panel including both NT-proBNP and neuropilin-1, generating a balanced accuracy of 0.81 with a sensitivity of 77.3% and a specificity of 86.5%. Therefore, while the addition of NT-proBNP or NT-proBNP plus neuropilin-1 slightly decreased the sensitivity and specificity in the derivation cohort, the addition of the two biomarkers improved the accuracy, sensitivity and specificity in the validation cohort.

Identifying biomarkers that predict clinical variables related to PAH

The combination of serological biomarkers and clinical variables to create composite scores has strengthened the conventional diagnostic approach in SSc patients [12]. Identifying additional biomarkers, beyond NT-proBNP and uric acid, that could predict clinical variables related to PAH would be highly advantageous and reduce the need for repeated invasive procedures, *e.g.* RHC. To investigate whether any of the protein biomarkers measured could accurately predict any of the recorded clinical variables we applied a SPLS regression analysis to the DETECT Discovery Cohort. Of the clinical variables tested (table 1), the association with our biomarker composite panel was generally weak, with PVR providing the best R^2 (R^2=0.321): NT-proBNP, RAGE, IGFBP-7, pyruvate carboxylase (cFib), vascular cell adhesion molecule (VCAM)-1 and surfactant protein D (SP-D) (figure 5). The highest correlation for PVR was obtained with NT-proBNP (r=0.46), RAGE (r=0.43) and IGFBP-7 (r=0.41). To verify the relevance of the identified variables we applied a RF analysis as an alternative to the SPLS analysis. RAGE, NT-proBNP, IGFBP-7, SP-D and VCAM-1 that were selected by SPLS to predict PVR were also among the 10 most important features according to a RF approach. When looking at these biomarkers in relation to all clinical variables, we observed that RAGE also showed a relevant correlation with FVC % pred (r=0.51, p=8.63e-12) (supplementary figure S10).

Discussion

PAH is a devastating complication of SSc, and there is evidence that early detection and treatment can improve the outcomes. The DETECT algorithm is a frequently used screening model for the detection of PAH in SSc patients [12]. It contains eight variables, including clinical variables from multiple tests and two circulating biomarkers, *i.e.* NT-proBNP and serum uric acid, both reflecting cardiac dysfunction [15–17]. The sensitivity for the detection of PAH using DETECT is high (96%), but the specificity is relatively low (48%). Much effort is therefore currently given to the discovery of diagnostic biomarkers for the accurate and noninvasive prediction of PAH in SSc patients and other "at-risk" populations.

In this study we used serum samples from the DETECT study and an unbiased high-throughput assay platform to discover novel protein biomarkers with the potential to aid screening and diagnosis. We have

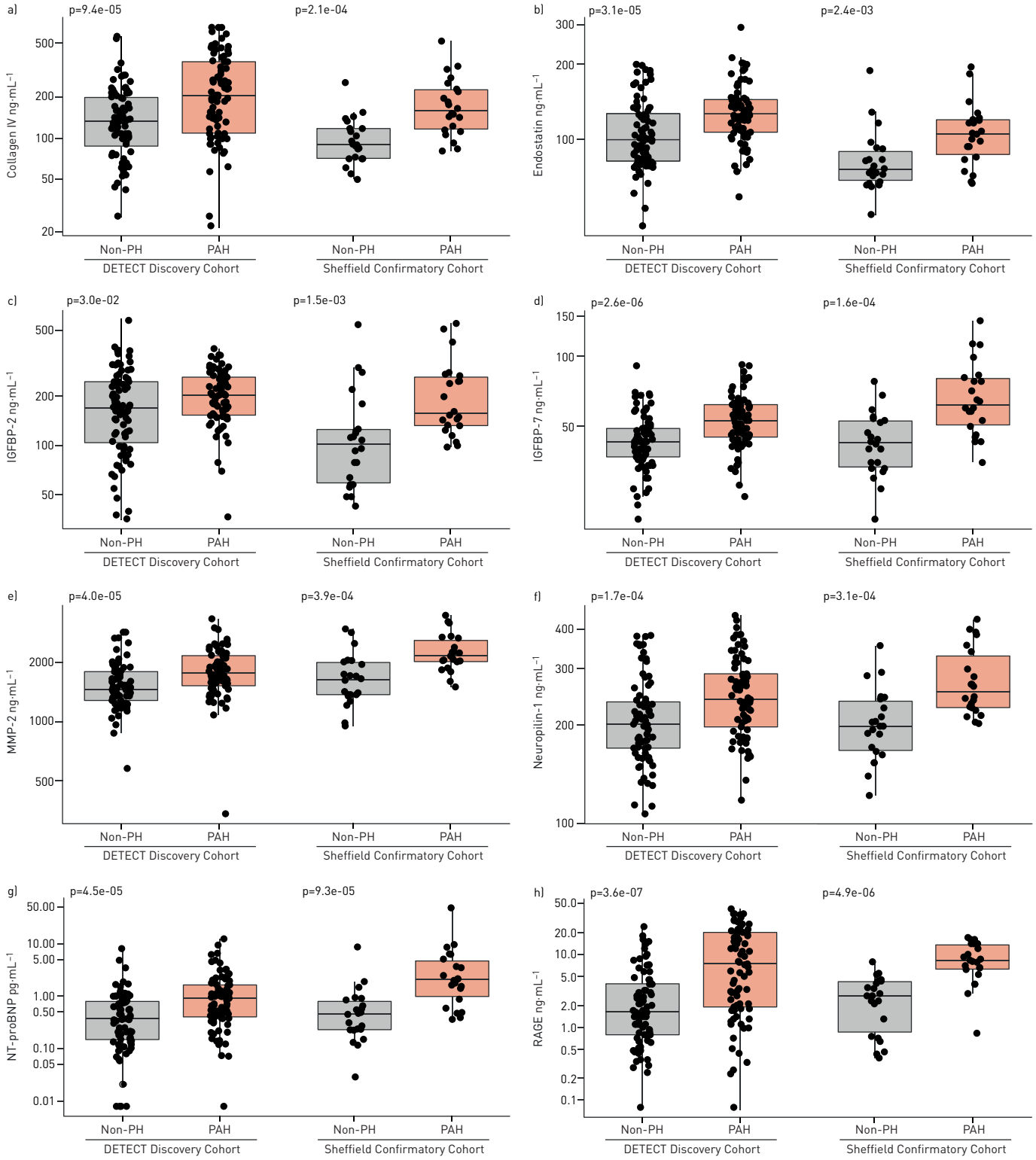

FIGURE 3 Serum concentrations of the eight best-performing and common proteins in predicting pulmonary arterial hypertension (PAH) in the DETECT Discovery Cohort and the Sheffield Confirmatory Cohort: a) collagen IV, b) endostatin, c) insulin-like growth factor binding protein (IGFBP)-2, d) IGFBP-7, e) matrix metallopeptidase (MMP)-2, f) neuropilin-1, g) N-terminal pro-brain natriuretic peptide (NT-proBNP) and h) RAGE (receptor for advanced glycation end products). PH: pulmonary hypertension. Boxes indicate median and interquartile range; whiskers indicate the full range of the data. Individual patient samples are represented by dots. p-values from the Wilcoxon rank-sum test between the two patient groups.

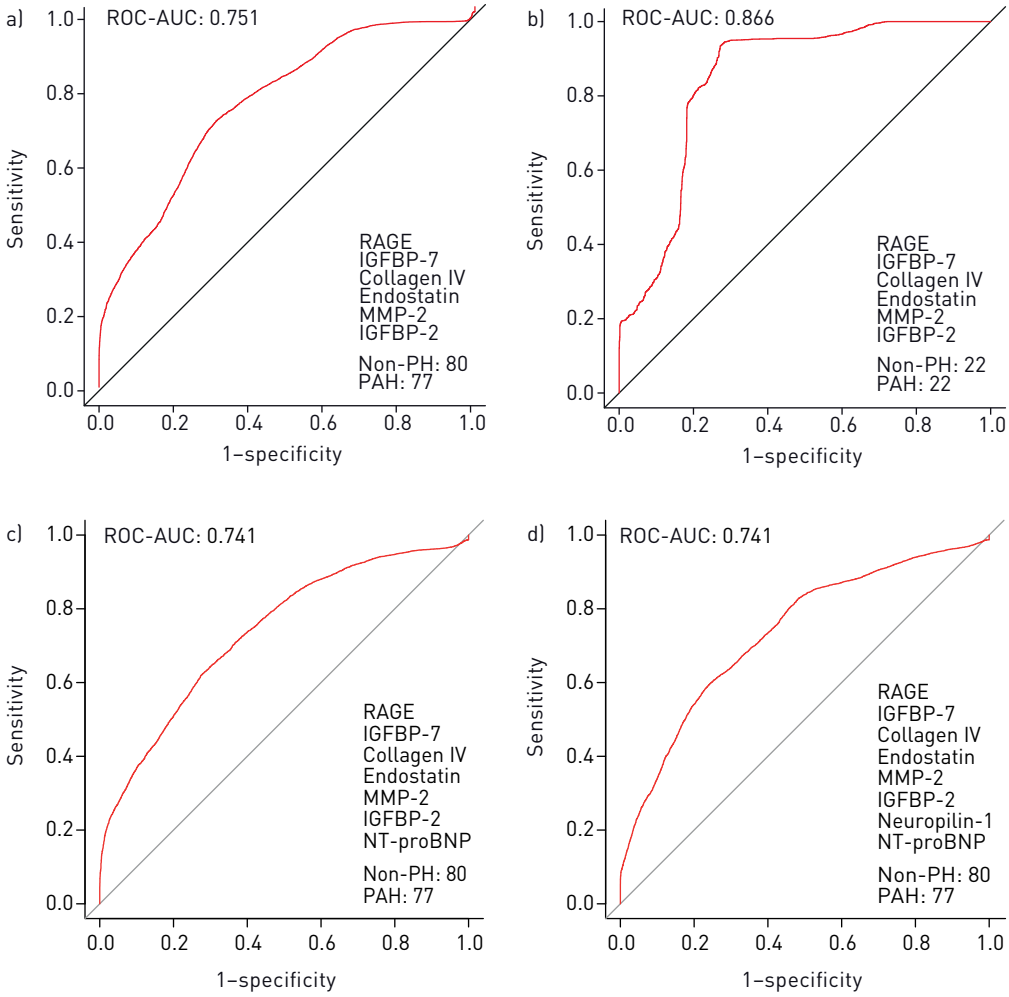

FIGURE 4 a, b) Performance of the panel of six common protein biomarkers in a) the DETECT Discovery Cohort and b) the Sheffield Confirmatory Cohort: receiver operating characteristic (ROC) curves of the pulmonary arterial hypertension (PAH) *versus* non-pulmonary hypertension (non-PH) classifier. ROC-AUC: area under the ROC curve; RAGE: receptor for advanced glycation end products; IGFBP: insulin-like growth factor binding protein; MMP: matrix metallopeptidase; SSc: systemic sclerosis; NT-proBNP: N-terminal pro-brain natriuretic peptide. The six selected proteins are the subset from the eight common proteins that produced the best ROC-AUC in the DETECT Discovery Cohort (0.751). c, d) Addition of c) NT-proBNP or d) NT-proBNP plus neuropilin-1 to the six selected proteins.

identified and validated a panel of eight biomarkers, *i.e.* RAGE, IGFBP-7, collagen IV, endostatin, MMP-2, IGFBP-2, NT-proBNP and neuropilin-1, with potential for classifying patients with PAH in a mixed population of patients with SSc.

Several of the proteins identified have been previously found to play significant roles in pulmonary vascular remodelling (RAGE and MMP-2), angiogenesis and cellular growth (collagen IV, endostatin, IGFBP-2 and neuropilin-1), and cardiac dysfunction (NT-proBNP and IGFBP-7). Among the top-ranking proteins, RAGE plays an important role in the accumulation of extracellular matrix proteins and particularly in vascular remodelling [18–23]. RAGE expression has been shown to be upregulated in pulmonary arteries that were isolated from Sugen 5416 plus hypoxia (SuHx)-induced PH mice and deletion of RAGE protects these mice from PH. Serum levels of soluble RAGE were also shown to be higher compared with controls in patients with idiopathic PAH and chronic thromboembolic pulmonary hypertension [24, 25] and in SuHx PH mice [22], and *in vitro* soluble RAGE has been shown to regulate bone morphogenetic proteins and calcium binding protein S100A4-induced proliferation and migration in pulmonary artery smooth muscle cells [20, 26]. The diagnostic value of RAGE was maintained when adjusted for age in our study (data not shown), negating a response to increasing advanced glycation products with age. We also found that RAGE levels were lower in serum of patients with diffuse compared with limited SSc, which excluded a relationship between RAGE expression levels and the extent of skin

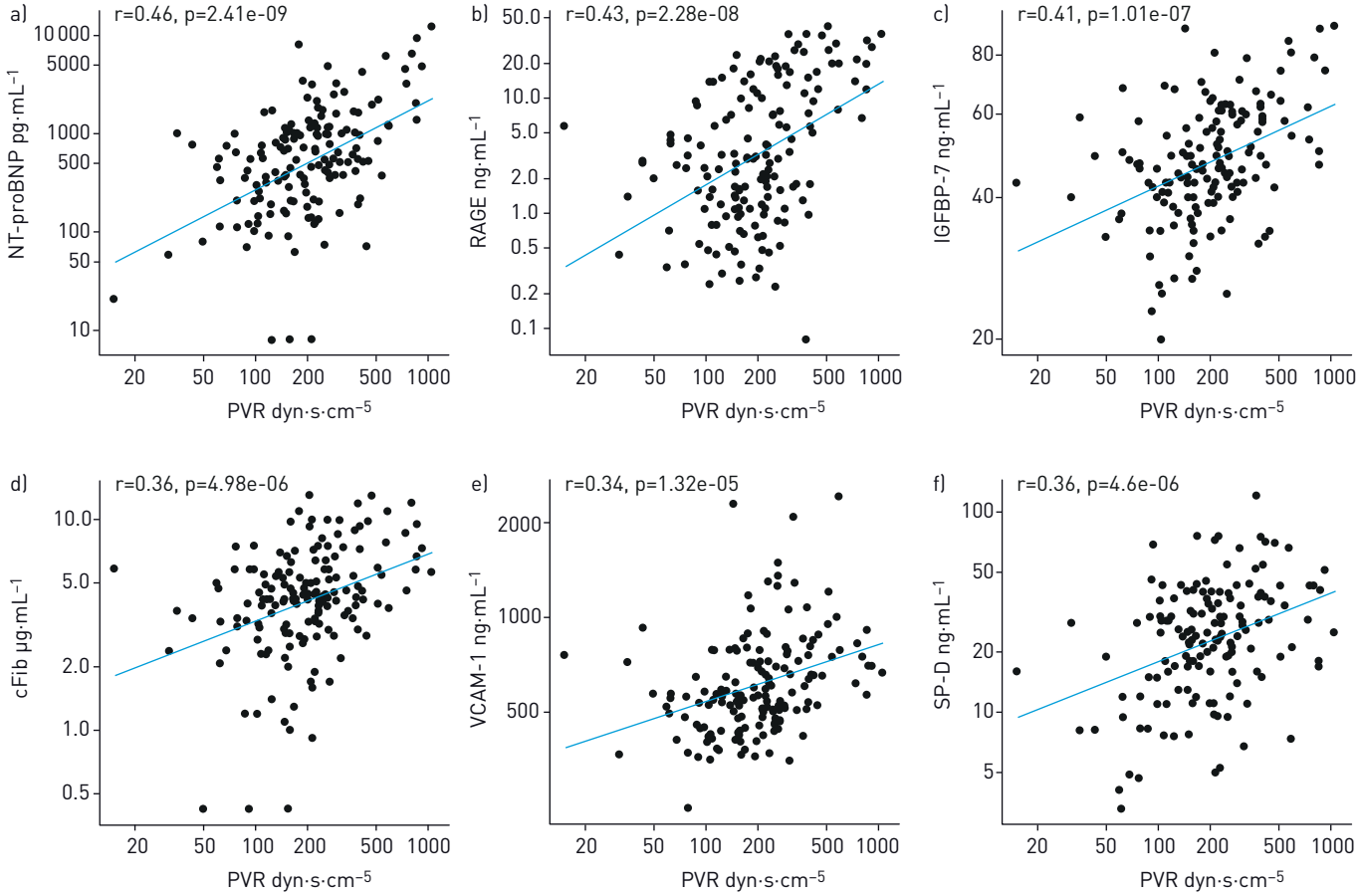

FIGURE 5 Sparse partial least squares association of pulmonary vascular resistance (PVR) to six common biomarker proteins: a) N-terminal pro-brain natriuretic peptide (NT-proBNP), b) RAGE (receptor for advanced glycation end products), c) insulin-like growth factor binding protein (IGFBP)-7, d) pyruvate carboxylase (cFib), e) vascular cell adhesion molecule (VCAM)-1 and f) surfactant protein D (SP-D). Correlation plots for each individual biomarker variable with PVR, showing Pearson's correlation coefficient between the logarithm of the two variables and the corresponding p-value.

and organ fibrosis involvement (supplementary figure S11). MMP-2, the other marker of vascular remodelling, is a metalloproteinase involved in the breakdown of the extracellular matrix and collagen IV, and contributes to the degradation of basal membranes [27]. Interestingly, hypoxia can attenuate the physiological postnatal increase of MMP-2 expression, which impacts alveolar development and associated pulmonary arterial remodelling [28]. It was recently shown that MMP-2 expression increased under hypoxia in pulmonary artery endothelium concomitant with a thickening of blood vessels. Inhibition of MMP-2 in a mouse model of PH prevented the development of PH and the proliferation of pulmonary artery endothelial cells under hypoxia [29].

IGFBP-7, the protein ranked second in our panel, has been associated with cellular senescence and cardiac dysfunction [30–33], and has the potential to complement the role of NT-proBNP, an established marker of cardiac stress [34]. Endostatin (collagen XVIII) and collagen IV, ranking at positions 3 and 4 in our panel, are important components of the extracellular vascular basement membrane that separates, for instance, epithelial cells and endothelial cells in the heart [35–37]. Endostatin was previously reported as a potential biomarker in PAH that could predict adverse outcome [38]. Although the role of collagen IV in PAH has not been described specifically, collagen IV synthesis can be promoted by nitric oxide production and collagen IV increase was shown to contribute to angiogenesis of lung endothelial cells [38]. Neuropilin-1, another molecule involved in angiogenesis, interacts with vascular endothelial growth factor (VEGF) family members to stimulate angiogenesis in endothelial cells [39, 40]. VEGF receptors are expressed by endothelial cells within the plexiform lesions in patients with PAH [41, 42]. These different markers of angiogenesis have therefore direct implication in the pathology of PAH and may relate to the early development of PAH in SSc patients. The potential role that these proteins could play in various aspects of PAH pathobiology provides encouragement that the proteins selected may have sensitivity to

disease-modifying therapies, although a further study on longitudinal samples would be required to demonstrate this.

The loss in performance of the six-protein panel derived from the DETECT Discovery Cohort when tested in the Sheffield Confirmatory Cohort most likely relates to disease heterogeneity in both SSc and PAH. Other contributing factors could be reflective of different sample processing or that the Sheffield Confirmatory Cohort was obtained from patients referred to a specialist PAH centre and they therefore had a higher probability of having PAH; indeed, the Sheffield Confirmatory Cohort may have also had more advanced PAH (supplementary figures S1–S9) as reflected by the 2-fold higher median NT-proBNP when compared with the DETECT Discovery Cohort. However, it is also encouraging that the protein panel performs well within both the general rheumatology setting where PAH may be less severe and the specialist PH centre with potentially more advanced disease. It is important than any biomarker/panel has sensitivity across the spectrum of disease. NT-proBNP is a widely used biomarker for PAH, reflecting an elevated right ventricle overload. However, raised NT-proBNP is not specific to PAH since other pathological conditions can lead to an increase in right ventricle overload and NT-proBNP levels [7].

As well as identifying a biomarker panel to assist early diagnosis of patients with PAH and SSc, we also examined whether we could identify potential protein biomarker surrogates for clinical variables commonly used to diagnose PAH. Among those tested, PVR was the clinical parameter best predicted by a biomarker panel. This panel, consisting of five proteins (RAGE, NT-proBNP, SP-D, VCAM-1 and cFib), contained two proteins (RAGE and NT-proBNP) that overlapped with the most reproducible eight-protein diagnostic panel (RAGE, IGFBP-7, collagen IV, endostatin, MMP-2, IGFBP-2, NT-proBNP and neuropilin-1). While cFib was not confirmed using a second analytical method (RF), all other markers were confirmed. SP-D has a well-recognised role in regulating inflammation and VCAM-1 has a well-recognised role in the mediation of leukocyte–endothelial cell adhesion, supporting the concept that the proteins may influence pulmonary vascular remodelling and therefore PVR. PAH is characterised by progressive pulmonary vascular remodelling leading to increased PVR, and eventually to right heart failure and death [43, 44]. Since PVR reflects the progressive pathogenesis of PAH [45], monitoring PVR in response to treatment by any method, but particularly a noninvasive method, is highly desirable. The use of this panel may therefore provide valuable information on ongoing pathogenic processes in SSc-PAH patients.

In this study, our biomarker panel identifies patients with SSc-PAH in a superior manner to NT-proBNP alone (ROC-AUC 0.689) (data not shown), perhaps suggesting that our panel detects early PAH before cardiac stress becomes dominant. It is interesting therefore to examine how this panel might perform under the new 2018 World Symposium on Pulmonary Hypertension recommended classification (mPAP 21–24 mmHg) to identify patients that have mild or early PAH [46]. However, this would have to be tested in larger relevant cohorts of patients. Since our initial cohort selection and analysis it has been recognised that patients with a borderline elevation of mPAP (*i.e.* mPAP 21–24 mmHg) can develop symptoms comparable to patients with mPAP $\geqslant$25 mmHg. Specifically, they have increased risk of progression to $\geqslant$25 mmHg and had a higher mortality rate than patients with mPAP <21 mmHg [47, 48]. Given the recent recommendation to change the cut-off for the definition of PH to include patients with mPAP >20 mmHg [46] we reran the analysis with this new criteria (mPAP >20 mmHg, PVR $\geqslant$3 WU and PAWP $\leqslant$15 mmHg). Reassuringly, only two biomarkers, *i.e.* IGFBP-7 and neuropilin-1 (the two weakest variables of importance), were dropped from the eight variables identified, suggesting that the remaining six proteins are particularly sensitive to early changes associated with PAH in the context of SSc.

A previous study by Rice *et al.* [49] identified both midkine and follistatin-like 3 as two proteins that might serve as SSc-PAH biomarkers, albeit in a small discovery cohort of only 13 patients, all with limited SSc. Our study included patients with diffuse, limited and mixed SSc in the DETECT Discovery Cohort, which is more reflective of the patient population within the rheumatology setting. Unfortunately, neither protein was within the Myriad DiscoveryMAP version 3.3 platform, so we were unable to determine whether these proteins could classify PAH in both of our cohorts. Future prospective studies should look to compare or incorporate these proteins.

A limitation of our current study is the lack of longitudinal data. Testing whether the protein panel changes in response to treatment and disease progression is an important next step. This is pertinent not only to PAH-specific therapies but also to background immunosuppressive therapies that many of these patients will be treated with. One other significant limitation of our study is the matched PAH and non-PH cohort size. With an estimated 10% transition of patients with SSc to develop PAH, future studies looking to validate these models would ideally have a more representative 1:10 PAH:no-PH proportion. We also acknowledge that our patient cohorts do not fully reflect the SSc patient population. The patients in the Sheffield Confirmatory Cohort were more advanced in their diagnosis process and had a high

suspicion of having PAH for referral to a specialist centre, and the improved performance of the biomarker panel trained on the DETECT Discovery Cohort may reflect this. However, this study provides important proof-of-concept data that applying machine learning tools to proteomic data can identify protein biomarkers to help screen patients at risk of PAH. Although not directly tested here, it is highly possible that this protein panel, developed on PAH in the context of SSc, may have utility in other forms of PAH or identify patients in other non-PAH diagnostic groups who have some pulmonary vascular remodelling.

The ultimate aim of this study is to identify a screening protein panel that can be incorporated into a future iteration of the DETECT algorithm to enhance the sensitivity and specificity of the current DETECT algorithm. Clearly, before this can be achieved the current protein panel will require further validation and "tuning" in a prospective and longitudinal SSc cohort within the rheumatology setting. Although challenges remain, integrating proteomic profiling into an existing screening programme such as DETECT should be achievable.

Acknowledgements: The authors would like to thank all investigators and patients involved in the DETECT study and those participating in The Sheffield Teaching Hospitals Observational Study of Patients with Pulmonary Hypertension, Cardiovascular and Lung Disease (STH-ObS).

Author contributions: All authors contributed to the study design, data acquisition, and analysis and interpretation of the data. Y. Bauer and A. Lawrie drafted the manuscript. All authors read and revised the content critically, approved the final manuscript, and are accountable for its content.

Conflict of interest: Y. Bauer is a former employee of Actelion Pharmaceuticals Ltd and Idorsia Pharmaceuticals Ltd, and is now an employee of Galapagos GmbH. S. de Bernard reports grants from Idorsia, during the conduct of the study. P. Hickey has nothing to disclose. K. Ballard is an employee of Myriad RBM. J. Cruz is an employee of Myriad RBM. P. Cornelisse is a former employee of Actelion Pharmaceuticals Ltd. H. Chadha-Boreham is a former employee of Actelion Pharmaceuticals Ltd. O. Distler reports personal fees for consultancy from Amgen, AbbVie, Acceleron Pharma, AnaMar, Actelion, Alexion, Arxx Therapeutics, Baecon Discovery, Blade Therapeutics, Corbuspharma, CSL Behring, ChemomAb, Horizon Pharmaceuticals, Ergonex, Galapagos NV, Glenmark Pharmaceuticals, GSK, Inventiva, Italfarmaco, iQone, iQvia, Kymera, Lilly, Medac, Sanofi, Target Bio Science and UCB, grants and personal fees for consultancy and lectures from Bayer and Boehringer Ingelheim, personal fees for interviewing from Catenion, grants from Competitive Drug Development International Ltd, personal fees for consultancy and lectures from Medscape, MSD, Pfizer and Roche, grants and personal fees for consultancy from Mitsubishi Tanabe Pharma, personal fees for lectures from Novartis, outside the submitted work; and has a patent mir-29 for the treatment of systemic sclerosis issued (US8247389, EP2331143). D. Rosenberg is an employee of and hold shares in Johnson and Johnson. M. Doelberg is an employee of Actelion Pharmaceuticals Ltd. S. Roux is a former employee of Actelion Pharmaceuticals Ltd. O. Nayler is a former employee and former stock owner of Actelion Pharmaceuticals Ltd, and current employee and stock owner of Idorsia Pharmaceuticals Ltd. A. Lawrie reports grants from the British Heart Foundation and Medical Research Council, grants, personal fees and other (conference attendance and travel) from Actelion Pharmaceuticals, grants and personal fees from GlaxoSmithKline, outside the submitted work.

Support statement: The work of Y. Bauer, O. Nayler and S. de Bernard was funded by Actelion Pharmaceuticals Ltd. P. Hickey was funded by a Donald Health Clinical Research Training Fellowship funded in partnership between Actelion Pharmaceuticals, Sheffield Teaching Hospitals Foundation NHS Trust and the University of Sheffield, and A. Lawrie was funded by British Heart Foundation (BHF) Senior Basic Science Research Fellowships (FS/13/48/30453 and FS/18/52/33808). Recruitment and collection of samples to The Sheffield Teaching Hospitals Observational Study of Patients with Pulmonary Hypertension, Cardiovascular and Lung Disease (STH-ObS) was supported by BHF PG/11/116/29288 and the Sheffield NIHR Clinical Research Facility. The views expressed in this manuscript are those of the authors and not necessarily those of Actelion Pharmaceuticals Ltd, Myriad RBM, the BHF, the NHS, the NIHR or the Dept of Health. Funding information for this article has been deposited with the Crossref Funder Registry.

References

1 Mukerjee D. Prevalence and outcome in systemic sclerosis associated pulmonary arterial hypertension: application of a registry approach. *Ann Rheum Dis* 2003; 62: 1088–1093.
2 Hachulla E, Gressin V, Guillevin L, *et al.* Early detection of pulmonary arterial hypertension in systemic sclerosis: a French nationwide prospective multicenter study. *Arthritis Rheum* 2005; 52: 3792–3800.
3 Tyndall AJ, Bannert B, Vonk M, *et al.* Causes and risk factors for death in systemic sclerosis: a study from the EULAR Scleroderma Trials and Research (EUSTAR) database. *Ann Rheum Dis* 2010; 69: 1809–1815.
4 Humbert M, Sitbon O, Chaouat A, *et al.* Pulmonary arterial hypertension in France: results from a national registry. *Am J Respir Crit Care Med* 2006; 173: 1023–1030.
5 Chung L, Domsic RT, Lingala B, *et al.* Survival and predictors of mortality in systemic sclerosis-associated pulmonary arterial hypertension: outcomes from the pulmonary hypertension assessment and recognition of outcomes in scleroderma registry. *Arthritis Care Res* 2014; 66: 489–495.
6 Hurdman J, Condliffe R, Elliot CA, *et al.* ASPIRE registry: Assessing the Spectrum of Pulmonary hypertension Identified at a REferral centre. *Eur Respir J* 2012; 39: 945–955.
7 Hickey PM, Lawrie A, Condliffe R. Circulating protein biomarkers in systemic sclerosis related pulmonary arterial hypertension: a review of published data. *Front Med* 2018; 5: 175.
8 Bergemann R, Allsopp J, Jenner H, *et al.* High levels of healthcare utilization prior to diagnosis in idiopathic pulmonary arterial hypertension support the feasibility of an early diagnosis algorithm: the SPHInX project. *Pulm Circ* 2018; 8: 2045894018798613.

9 Kiely DG, Doyle O, Drage E, *et al.* Utilising artificial intelligence to determine patients at risk of a rare disease: idiopathic pulmonary arterial hypertension. *Pulm Circ* 2019; 9: 2045894019890549.
10 Kiely DG, Lawrie A, Humbert M. Screening strategies for pulmonary arterial hypertension. *Eur Heart J Suppl* 2019; 21: K9–K20.
11 Humbert M, Yaici A, de Groote P, *et al.* Screening for pulmonary arterial hypertension in patients with systemic sclerosis: clinical characteristics at diagnosis and long-term survival. *Arthritis Rheum* 2011; 63: 3522–3530.
12 Coghlan JG, Denton CP, Grünig E, *et al.* Evidence-based detection of pulmonary arterial hypertension in systemic sclerosis: the DETECT study. *Ann Rheum Dis* 2014; 73: 1340–1349.
13 Hao Y, Thakkar V, Stevens W, *et al.* A comparison of the predictive accuracy of three screening models for pulmonary arterial hypertension in systemic sclerosis. *Arthritis Res Ther* 2015; 17: 7.
14 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.
15 Gigante A, Barbano B, Barilaro G, *et al.* Serum uric acid as a marker of microvascular damage in systemic sclerosis patients. *Microvasc Res* 2016; 106: 39–43.
16 Maiuolo J, Oppedisano F, Gratteri S, *et al.* Regulation of uric acid metabolism and excretion. *Int J Cardiol* 2016; 213: 8–14.
17 Odler B, Foris V, Gungl A, *et al.* Biomarkers for pulmonary vascular remodeling in systemic sclerosis: a pathophysiological approach. *Front Physiol* 2018; 9: 587.
18 Sakaguchi T, Yan SF, Yan SD, *et al.* Central role of RAGE-dependent neointimal expansion in arterial restenosis. *J Clin Invest* 2003; 111: 959–972.
19 Spiekerkoetter E, Lawrie A, Merklinger S, *et al.* Mts1/S100A4 stimulates human pulmonary artery smooth muscle cell migration through multiple signaling pathways. *Chest* 2005; 128: 577S.
20 Spiekerkoetter E, Guignabert C, de Jesus Perez V, *et al.* S100A4 and bone morphogenetic protein-2 codependently induce vascular smooth muscle cell migration *via* phospho-extracellular signal-regulated kinase and chloride intracellular channel 4. *Circ Res* 2009; 105: 639–647.
21 Nakamura K, Sakaguchi M, Matsubara H, *et al.* Crucial role of RAGE in inappropriate increase of smooth muscle cells from patients with pulmonary arterial hypertension. *PLoS One* 2018; 13: e0203046.
22 Jia D, He Y, Zhu Q, *et al.* RAGE-mediated extracellular matrix proteins accumulation exacerbates HySu-induced pulmonary hypertension. *Cardiovasc Res* 2017; 113: 586–597.
23 Sakaguchi M, Kinoshita R, Putranto EW, *et al.* Signal diversity of receptor for advanced glycation end products. *Acta Med Okayama* 2017; 71: 459–465.
24 Moser B, Megerle A, Bekos C, *et al.* Local and systemic RAGE axis changes in pulmonary hypertension: CTEPH and iPAH. *PLoS One* 2014; 9: e106440.
25 Suzuki S, Nakazato K, Sugimoto K, *et al.* Plasma levels of receptor for advanced glycation end-products and high-mobility group box 1 in patients with pulmonary hypertension. *Int Heart J* 2016; 57: 234–240.
26 Lawrie A, Spiekerkoetter E, Martinez EC, *et al.* Interdependent serotonin transporter and receptor pathways regulate S100A4/Mts1, a gene associated with pulmonary vascular disease. *Circ Res* 2005; 97: 227–235.
27 Wang H, Su Y. Collagen IV contributes to nitric oxide-induced angiogenesis of lung endothelial cells. *Am J Physiol Cell Physiol* 2011; 300: C979–C988.
28 Ambalavanan N, Nicola T, Li P, *et al.* Role of matrix metalloproteinase-2 in newborn mouse lungs under hypoxic conditions. *Pediatr Res* 2008; 63: 26–32.
29 Liu Y, Zhang H, Yan L, *et al.* MMP-2 and MMP-9 contribute to the angiogenic effect produced by hypoxia/15-HETE in pulmonary endothelial cells. *J Mol Cell Cardiol* 2018; 121: 36–50.
30 Januzzi JL, Packer M, Claggett B, *et al.* IGFBP7 (insulin-like growth factor-binding protein-7) and neprilysin inhibition in patients with heart failure. *Circ Heart Fail* 2018; 11: e005133.
31 Gouda P, Ezekowitz JA. Update on the diagnosis and management of acute heart failure. *Curr Opin Cardiol* 2019; 34: 202–206.
32 Gandhi PU, Gaggin HK, Redfield MM, *et al.* Insulin-like growth factor-binding protein-7 as a biomarker of diastolic dysfunction and functional capacity in heart failure with preserved ejection fraction: results from the RELAX trial. *JACC Heart Fail* 2016: 860–869.
33 Barroso MC, Kramer F, Greene SJ, *et al.* Serum insulin-like growth factor-1 and its binding protein-7: potential novel biomarkers for heart failure with preserved ejection fraction. *BMC Cardiovasc Disord* 2016; 16: 199.
34 Wang TJ, Larson MG, Levy D, *et al.* Plasma natriuretic peptide levels and the risk of cardiovascular events and death. *N Engl J Med* 2004; 350: 655–663.
35 Tomono Y, Naito I, Ando K, *et al.* Epitope-defined monoclonal antibodies against multiplexin collagens demonstrate that type XV and XVIII collagens are expressed in specialized basement membranes. *Cell Struct Funct* 2002; 27: 9–20.
36 Ortega N. New functional roles for non-collagenous domains of basement membrane collagens. *J Cell Sci* 2002: 4201–4214.
37 Marchand M, Monnot C, Muller L, *et al.* Extracellular matrix scaffolding in angiogenesis and capillary homeostasis. *Semin Cell Dev Biol* 2019; 89: 147–156.
38 Damico R, Kolb TM, Valera L, *et al.* Serum endostatin is a genetically determined predictor of survival in pulmonary arterial hypertension. *Am J Respir Crit Care Med* 2015; 191: 208–218.
39 Finney AC, Orr AW. Guidance molecules in vascular smooth muscle. *Front Physiol* 2018; 9: 1311.
40 Wang L, Zeng H, Wang P, *et al.* Neuropilin-1-mediated vascular permeability factor/vascular endothelial growth factor-dependent endothelial cell migration. *J Biol Chem* 2003; 278: 48848–48860.
41 Jonigk D, Golpon H, Bockmeyer CL, *et al.* Plexiform lesions in pulmonary arterial hypertension composition, architecture, and microenvironment. *Am J Pathol* 2011; 179: 167–179.
42 Tuder RM, Chacon M, Alger L, *et al.* Expression of angiogenesis-related molecules in plexiform lesions in severe pulmonary hypertension: evidence for a process of disordered angiogenesis. *J Pathol* 2001; 195: 367–374.

43 Chan SY, Loscalzo J. Pathogenic mechanisms of pulmonary arterial hypertension. *J Mol Cell Cardiol* 2008; 44: 14–30.
44 Humbert M, Sitbon O, Simonneau G. Treatment of pulmonary arterial hypertension. *N Engl J Med* 2004; 351: 1425–1436.
45 Stenmark KR, Fagan KA, Frid MG. Hypoxia-induced pulmonary vascular remodeling: cellular and molecular mechanisms. *Circ Res* 2006; 99: 675–691.
46 Simonneau G, Montani D, Celermajer DS, *et al.* Haemodynamic definitions and updated clinical classification of pulmonary hypertension. *Eur Respir J* 2019; 53: 1801913.
47 Visovatti SH, Distler O, Coghlan JG, *et al.* Borderline pulmonary arterial pressure in systemic sclerosis patients: a post-hoc analysis of the DETECT study. *Arthritis Res Ther* 2014; 16: 493.
48 Hoeper MM, Humbert M. The new haemodynamic definition of pulmonary hypertension: evidence prevails, finally! *Eur Respir J* 2019; 53: 1900038.
49 Rice LM, Mantero JC, Stratton EA, *et al.* Serum biomarker for diagnostic evaluation of pulmonary arterial hypertension in systemic sclerosis. *Arthritis Res Ther* 2018; 20: 185.