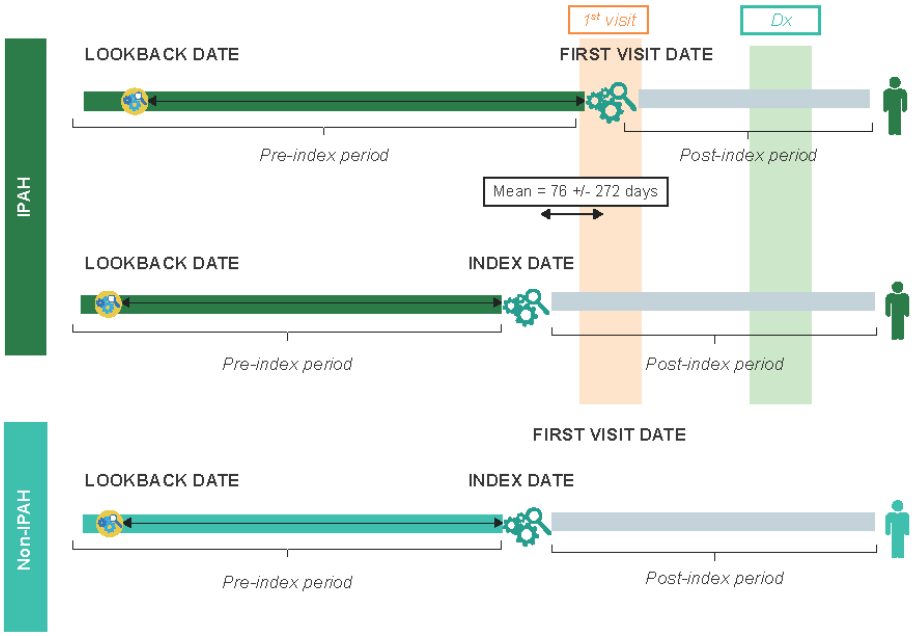
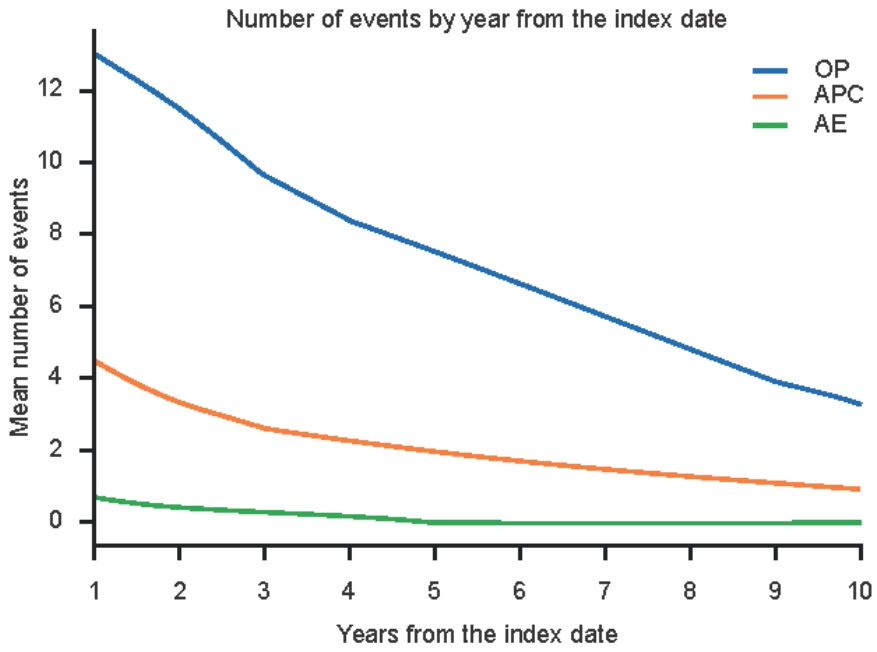


Supplementary Figure 1: Graphical representation of the patient history window considered in the predictive model for each cohort.

Dx, diagnosis; iPAH, idiopathic pulmonary arterial hypertension



Supplementary Figure 2: Mean number of events by year prior to the index date, for each HES table, for the iPAH cohort. APC table = inpatient records; OP table = outpatient records; A&E table = accident and emergency records.

HES, Hospital Episode Statistics; iPAH, idiopathic pulmonary arterial hypertension.

Supplementary Table 1: ICD-10 codes for the definition of the non-iPAH cohort dataset

	Detail
I420	Dilated cardiomyopathy
E039	Hypothyroidism, unspecified
I050	Rheumatic mitral stenosis
I342	Nonrheumatic mitral (valve) stenosis
Q232	Congenital mitral stenosis
M351	Other overlap syndromes
G473	Sleep apnoea
M32	Systemic lupus erythematosus (SLE)
K766	Portal hypertension
I370	Nonrheumatic pulmonary valve stenosis
L940	Localized scleroderma [morphea]
L941	Linear scleroderma
M43	Spondylolysis
I20	Angina pectoris
I21	ST elevation (STEMI) and non-ST elevation (NSTEMI) myocardial infarction
I22	Subsequent ST elevation (STEMI) and non-ST elevation (NSTEMI) myocardial infarction
I23	Certain current complications following ST elevation (STEMI) and non-ST elevation (NSTEMI) myocardial infarction (within the 28 day period)
I24	Other acute ischaemic heart diseases
I25	Chronic ischaemic heart disease
I26	Pulmonary embolism

I27	Other pulmonary heart diseases
I28	Other diseases of pulmonary vessels
I50	Heart failure
J40	Bronchitis, not specified as acute or chronic
J41	Simple and mucopurulent chronic bronchitis
J42	Unspecified chronic bronchitis
J43	Emphysema
J44	Other chronic obstructive pulmonary disease
J45	Asthma
J47	Bronchiectasis
J849	Interstitial pulmonary disease, unspecified

ICD-10, International Statistical Classification of Diseases and Related Health Problems, 10th Revision; iPAH, idiopathic pulmonary arterial hypertension.

Supplementary Table 2: Baseline characteristics of patients with and without confirmed diagnosis of idiopathic pulmonary arterial hypertension.

	All patients (n=709)	Patients with iPAH < 50 years (n=172)	Patients with iPAH ≥ 50 years (n=537)	Patients without iPAH (n=2,812,458)
Demographics				
Age (years)	60 (23)	38 (16)	66 (13)	71 (23)
Gender (% female)	62	70	60	51
Comorbidities				
Systemic hypertension (%)	48	20	59	60
Atrial fibrillation / flutter (%)	17	4	22	37
Ischaemic heart disease (%)	8	*	11	18

*Due to NHS Digital's reporting policy of small numbers (n<5), the percentage of patients by age group has been masked to main patient confidentiality; Values are median (interquartile range) or percentage, unless otherwise indicated. Age is reported at index date; co-morbidities are summarised over the observation window prior to the index date.

iPAH, idiopathic pulmonary arterial hypertension; NHS, National Health Service.

Supplementary Table 3: Baseline characteristics of patients with idiopathic pulmonary arterial hypertension.

	All patients (n=709)	Patients < 50 years (n=172)	Patients ≥ 50 years (n=537)
Demographics			
Age (years)	60 (17)	36 (9)	68 (9)
Gender (%, Female)	62	70	60
Right heart catheter			
RAP (mmHg)	12 (6)	12 (7)	12 (6)
mPAP(mmHg)	53 (12)	57 (13)	51 (11)
CI (L/min/m2)	2.4 (0.8)	2.6 (0.9)	2.3 (0.7)
PCWP (mmHg)	12 (5)	11 (4)	13 (5)
PVR (Wood Units)	833 (399)	914 (438)	810 (385)
SvO_2 (%)	60 (9)	62 (10)	60 (8)
Lung function and exercise capacity			
FEV_1 (absolute)	1.9 (0.7)	2.4 (0.8)	1.7 (0.6)
FVC (absolute)	2.7 (1.0)	3.2 (1.0)	2.6 (0.9)
FEV_1/FVC ratio	0.94 (0.24)	0.95 (0.22)	0.94 (0.24)
DL_{CO} (absolute)	3.5 (2.2)	5.7 (2.3)	2.9 (1.7)
ISWD (m)	166 (177)	293 (205)	129 (149)
Comorbidities			
Systemic hypertension (%)	54	18	65

Atrial fibrillation/flutter (%)	21	4	26
Ischaemic heart disease (%)	11	*	>10*
HCRU events in period up to 3 years pre-diagnosis (n)	18.7	14.4	19.7

CI, cardiac index; DL_{CO} , diffusing capacity of lung for carbon monoxide; FEV_1 , forced expiratory volume in one second; FVC, forced vital capacity; ICD-10, International Statistical Classification of Diseases and Related Health Problems, 10th Revision; iPAH, idiopathic pulmonary arterial hypertension; ISWD, incremental shuttle walk distance; mPAP , mean pulmonary arterial pressure; n, number of patients; PCWP , pulmonary capillary wedge pressure; PVR, pulmonary vascular resistance; RAP, right atrial pressure; SvO_2 , mixed venous oxygen saturation.

Data from right heart catheterisation within 3 months are presented; Lung function data and exercise capacity as assessed by a formal test are presented within 3 months of diagnosis; Comorbidities identified from ICD-10 codes are presented; Index date is defined as the first event for which a patient is referred to a cardiology, respiratory medicine or neurology specialist in Sheffield; Due to National Health Service Digital's reporting policy of small numbers (n<5), the percentage of patients by age group has been masked to main patient confidentiality (represented by *); Values are mean (standard deviation) or percentage, unless otherwise indicated. Detailed diagnostic characteristics of iPAH Patients were included in the descriptive analysis but not in the machine learning algorithm.

Supplementary Table 4: Codes and diagnoses.

Code	HES field	Description
D751	Primary ICD-10	D751: Secondary polycythaemia
I270	Primary ICD-10	I270: Primary pulmonary hypertension
I272	Primary ICD-10	I272: Other secondary pulmonary hypertension
I279	Primary ICD-10	I279: Pulmonary heart disease unspecified
I471	Primary ICD-10	I471: Supraventricular tachycardia
I48X	Primary ICD-10	I48X: Atrial fibrillation and flutter
I500	Primary ICD-10	I500: Congestive heart failure
I509	Primary ICD-10	I509: Heart failure unspecified
I517	Primary ICD-10	I517: Cardiomegaly
J459	Primary ICD-10	J459: Asthma unspecified
J849	Primary ICD-10	J849: Interstitial pulmonary disease unspecified
J90X	Primary ICD-10	J90X: Pleural effusion not elsewhere classified
M349	Primary ICD-10	M349: Systemic sclerosis unspecified
Q211	Primary ICD-10	Q211: Atrial septal defect
R05X	Primary ICD-10	R05X: Cough
R060	Primary ICD-10	R060: Dyspnoea
R068	Primary ICD-10	R068: Other and unspecified abnormalities of breathing
R072	Primary ICD-10	R072: Precordial pain
R073	Primary ICD-10	R073: Other chest pain
R074	Primary ICD-10	R074: Chest pain unspecified
R42X	Primary ICD-10	R42X: Dizziness and giddiness
R55X	Primary ICD-10	R55X: Syncope and collapse

R91X	Primary ICD-10	R91X: Abnormal findings on diagnostic imaging of lung
U082	Primary OCPS codes	U082: Ultrasound of abdomen
U106	Primary OCPS codes	U106: Myocardial perfusion scan
U112	Primary OCPS codes	U112: Doppler ultrasound of vessels of extremities
E921	Primary OPCS codes	E921: Carbon monoxide transfer factor test
E924	Primary OPCS codes	E924: Blood gas analysis
E925	Primary OPCS codes	E925: Complex lung function exercise test
E926	Primary OPCS codes	E926: Simple lung function exercise test
E931	Primary OPCS codes	E931: Measurement of peak expiratory flow rate
E932	Primary OPCS codes	E932: Spirometry
E935	Primary OPCS codes	E935: Measurement of static lung volume
K631	Primary OPCS codes	K631: Angiocardiography of combination of right and left side of heart
K652	Primary OPCS codes	K652: Catheterisation of right side of heart NEC
L133	Primary OPCS codes	L133: Arteriography of pulmonary artery
U071	Primary OPCS codes	U071: Computed tomography of chest
U103	Primary OPCS codes	U103: Cardiac magnetic resonance imaging
U151	Primary OPCS codes	U151: Lung perfusion scanning NEC
U192	Primary OPCS codes	U192: 24 hour ambulatory electrocardiography
U194	Primary OPCS codes	U194: Exercise electrocardiography
U198	Primary OPCS codes	U198: Other specified diagnostic electrocardiography
U199	Primary OPCS codes	U199: Unspecified diagnostic electrocardiography
U201	Primary OPCS codes	U201: Transthoracic echocardiography
U202	Primary OPCS codes	U202: Transoesophageal echocardiography
U205	Primary OPCS codes	U205: Stress echocardiography

U208	Primary OPCS codes	U208: Other specified diagnostic echocardiography
U209	Primary OPCS codes	U209: Unspecified diagnostic echocardiography
U354	Primary OPCS codes	U354: Computed tomography of pulmonary arteries
X821	Primary OPCS codes	X821: Primary pulmonary hypertension drugs band 1
B965	Secondary ICD-10	B965: *Pseudomonas* (*aeruginosa*) as the cause of diseases classified to other chapters
D508	Secondary ICD-10	D508: Other iron deficiency anaemias
E039	Secondary ICD-10	E039: Hypothyroidism unspecified
E059	Secondary ICD-10	E059: Thyrotoxicosis unspecified
E877	Secondary ICD-10	E877: Fluid overload
F101	Secondary ICD-10	F101: Mental and behavioural disorders due to use of alcohol harmful use
F102	Secondary ICD-10	F102: Mental and behavioural disorders due to use of alcohol dependence syndrome
F112	Secondary ICD-10	F112: Mental and behavioural disorders due to use of opioids dependence syndrome
F171	Secondary ICD-10	F171: Mental and behavioural disorders due to use of tobacco harmful use
F172	Secondary ICD-10	F172: Mental and behavioural disorders due to use of tobacco dependence syndrome
F329	Secondary ICD-10	F329: Depressive episode unspecified
F412	Secondary ICD-10	F412: Mixed anxiety and depressive disorder
F419	Secondary ICD-10	F419: Anxiety disorder unspecified
G473	Secondary ICD-10	G473: Sleep apnoea
H360	Secondary ICD-10	H360: Diabetic retinopathy

I071	Secondary ICD-10	I071: Tricuspid insufficiency
I10X	Secondary ICD-10	I10X: Essential (primary) hypertension
I258	Secondary ICD-10	I258: Other forms of chronic ischaemic heart disease
I270	Secondary ICD-10	I270: Primary pulmonary hypertension
I272	Secondary ICD-10	I272: Other secondary pulmonary hypertension
I279	Secondary ICD-10	I279: Pulmonary heart disease unspecified
I313	Secondary ICD-10	I313: Pericardial effusion (noninflammatory)
I351	Secondary ICD-10	I351: Aortic (valve) insufficiency
I361	Secondary ICD-10	I361: Nonrheumatic tricuspid (valve) insufficiency
I451	Secondary ICD-10	I451: Other and unspecified right bundle-branch block
I471	Secondary ICD-10	I471: Supraventricular tachycardia
I48X	Secondary ICD-10	I48X: Atrial fibrillation and flutter
I500	Secondary ICD-10	I500: Congestive heart failure
I509	Secondary ICD-10	I509: Heart failure unspecified
I517	Secondary ICD-10	I517: Cardiomegaly
I518	Secondary ICD-10	I518: Other ill-defined heart diseases
I730	Secondary ICD-10	I730: Raynaud syndrome
I839	Secondary ICD-10	I839: Varicose veins of lower extremities without ulcer or inflammation
J439	Secondary ICD-10	J439: Emphysema unspecified
J440	Secondary ICD-10	J440: Chronic obstructive pulmonary disease with acute lower respiratory infection
J441	Secondary ICD-10	J441: Chronic obstructive pulmonary disease with acute exacerbation unspecified
J449	Secondary ICD-10	J449: Chronic obstructive pulmonary disease unspecified

J459	Secondary ICD-10	J459: Asthma unspecified
J47X	Secondary ICD-10	J47X: Bronchiectasis
J841	Secondary ICD-10	J841: Other interstitial pulmonary diseases with fibrosis
J849	Secondary ICD-10	J849: Interstitial pulmonary disease unspecified
J960	Secondary ICD-10	J960: Acute respiratory failure
K746	Secondary ICD-10	K746: Other and unspecified cirrhosis of liver
K766	Secondary ICD-10	K766: Portal hypertension
K920	Secondary ICD-10	K920: Haematemesis
K921	Secondary ICD-10	K921: Melaena
K922	Secondary ICD-10	K922: Gastrointestinal haemorrhage unspecified
L891	Secondary ICD-10	L891: Stage II decubitus ulcer
M329	Secondary ICD-10	M329: Systemic lupus erythematosus unspecified
M341	Secondary ICD-10	M341: CR(E)ST syndrome
M349	Secondary ICD-10	M349: Systemic sclerosis unspecified
M350	Secondary ICD-10	M350: Sicca syndrome [Sjögren]
M359	Secondary ICD-10	M359: Systemic involvement of connective tissue unspecified
Q210	Secondary ICD-10	Q210: Ventricular septal defect
Q211	Secondary ICD-10	Q211: Atrial septal defect
Q218	Secondary ICD-10	Q218: Other congenital malformations of cardiac septa
Q249	Secondary ICD-10	Q249: Congenital malformation of heart unspecified
R000	Secondary ICD-10	R000: Tachycardia unspecified
R002	Secondary ICD-10	R002: Palpitations
R060	Secondary ICD-10	R060: Dyspnoea
R068	Secondary ICD-10	R068: Other and unspecified abnormalities of breathing

R073	Secondary ICD-10	R073: Other chest pain
R074	Secondary ICD-10	R074: Chest pain unspecified
R18X	Secondary ICD-10	R18X: Ascites
R42X	Secondary ICD-10	R42X: Dizziness and giddiness
R53X	Secondary ICD-10	R53X: Malaise and fatigue
R55X	Secondary ICD-10	R55X: Syncope and collapse
R609	Secondary ICD-10	R609: Oedema unspecified
R91X	Secondary ICD-10	R91X: Abnormal findings on diagnostic imaging of lung
Z136	Secondary ICD-10	Z136: Special screening examination for cardiovascular disorders
Z867	Secondary ICD-10	Z867: Personal history of diseases of the circulatory system
Z870	Secondary ICD-10	Z870: Personal history of diseases of the respiratory system
Z877	Secondary ICD-10	Z877: Personal history of congenital malformations deformations and chromosomal abnormalities
Z922	Secondary ICD-10	Z922: Personal history of long-term (current) use of other medicaments
Z991	Secondary ICD-10	Z991: Dependence on respirator
E921	Secondary OPCS codes	E921: Carbon monoxide transfer factor test
K633	Secondary OPCS codes	K633: Angiocardiography of left side of heart NEC
K634	Secondary OPCS codes	K634: Coronary arteriography using two catheters

	Secondary OPCS	
K636	codes	K636: Coronary arteriography NEC
	Secondary OPCS	
K652	codes	K652: Catheterisation of right side of heart NEC
	Secondary OPCS	
K661	codes	K661: Cardiotachygraphy
	Secondary OPCS	
L133	codes	L133: Arteriography of pulmonary artery
	Secondary OPCS	
U103	codes	U103: Cardiac magnetic resonance imaging
	Secondary OPCS	
U117	codes	U117: Magnetic resonance angiography
	Secondary OPCS	
U151	codes	U151: Lung perfusion scanning NEC
	Secondary OPCS	
U199	codes	U199: Unspecified diagnostic electrocardiography
	Secondary OPCS	
U201	codes	U201: Transthoracic echocardiography
	Secondary OPCS	
U202	codes	U202: Transoesophageal echocardiography
	Secondary OPCS	
U354	codes	U354: Computed tomography of pulmonary arteries
	Secondary OPCS	
X821	codes	X821: Primary pulmonary hypertension drugs band 1

	Secondary OPCS	
X822	codes	X822: Primary pulmonary hypertension drugs band 2
	Secondary OPCS	
X823	codes	X823: Primary pulmonary hypertension drugs band 3
	Secondary OPCS	
Y731	codes	Y731: Cardiopulmonary bypass
	Secondary OPCS	
Z401	codes	Z401: Pulmonary artery

CREST, calcinosis, Raynaud's phenomenon, esophageal dysmotility, sclerodactyly, and telangiectasia; ICD-10, the International Statistical Classification of Diseases and Related Health Problems.

Supplementary Table 5: Clinical specialty codes included within the model

Code	Description
340	Respiratory Med
320	Cardiology
300	G Medicine
100	General Surgery
130	Ophthalmology
110	T&O
303	Clin Haem
120	ENT
301	Gastroenterology
	RESPIRATORY
341	PHYSIOLOGY
430	Geriatric Med
502	Gynaecology
101	Urology
330	Dermatology
400	Neurology
	DIAGNOSTIC
812	IMAGING
	Anticoagulant
324	Service
140	Oral Surgery

	Cardiothoracic
170	Surgery
180	A&E
410	Rheumatology
361	Nephrology
307	Diabetic Medicine
107	Vascular Surgery
650	Physiotherapy
302	Endocrinology
103	Breast Surgery
104	Colorectal Surgery
304	Clin Physi
190	Anaesthetics
191	Pain Mgmt
160	Plastic Surgery
710	Adult Mental Illness
800	Clin Oncology
173	Thoracic Surgery
501	Obstetrics
150	Neurosurgery
653	Podiatry
306	Hepatology
172	Cardiac Surgery

	Interventional
811	Radiology
141	Restorative Dent
	Maxillo-Facial
144	Surgery
314	Rehab
350	Inf Diseases
	Critical Care
192	Medicine
822	Chem Pathology
321	Paed Cardiology
420	Paediatrics
	Gynaecological
503	Oncology
654	Dietetics
655	Orthoptics

A&E, accident and emergency; ENT, ear, nose, throat.

Supplementary Table 6: Profiles of top scored patients at risk of iPAH. The profiles are presented in terms of the proportion of patients observed to have a particular clinical event across patients predicted to by iPAH in terms of true positives (top 100) and false positives (top 500).

	Top 100 iPAH (true positives) (%)	**Top 500 non-iPAH (false positives) (%)**
Primary Diagnoses Variables		
Count of primary ICD codes within 12 months	Suppressed	0.082
Count of primary ICD codes 12 months prior	Suppressed	0.02
R068: Other and unspecific abnormalities of breathing	Suppressed	Suppressed
R060: Dyspnoea	Suppressed	0.05
I279: Pulmonary heart disease unspecified	Suppressed	Suppressed
R074: Chest pain unspecified	Suppressed	0.026
J459: Asthma unspecified	Suppressed	Suppressed
I500: Congestive heart failure	Suppressed	0.032
I48X: Atrial fibrillation and flutter	Suppressed	Suppressed
Secondary diagnoses variables		
I071: Tricuspid insufficiency	0.08	0.134
I10X: Essential (primary) hypertension	0.16	0.332
I517: Cardiomegaly	0.13	0.236
I500: Congestive heart failure	0.07	0.178

Count of secondary ICD codes within 12 months	0.54	0.776
I48X: Atrial fibrillation and flutter	Suppressed	0.084
F171: Mental and behavioural disorders due to use of tobacco harmful use	Suppressed	0.03
Z867: Personal history of diseases of the circulatory system	Suppressed	0.142
J459: Asthma unspecified	0.08	0.076
Count of secondary ICD codes 12 months prior	0.11	0.354
J449: Chronic obstructive pulmonary disease unspecified	Suppressed	0.124
R060: Dyspnoea	0.06	0.058
I361: Nonrheumatic tricuspid (value) insufficiency	Suppressed	0.074
G473: Sleep apnea	Suppressed	0.048
J841: Other interstitial pulmonary diseases with fibrosis	Suppressed	0.122
Speciality visits		
Respiratory med	0.44	0.68
Cardiology	0.77	0.75
G Medicine	0.58	0.686
Clin haem	0.16	0.204
Count of speciality visits within 12 months	1	0.998
A&E	Suppressed	0.042
Geriatric Med	00.06	0.09
Count of speciality visits within 12 months prior	0.68	0.636
Gastroenterology	0.06	0.156

Gynaecology	0.21	0.084
General surgery	0.18	0.174
Respiratory physiology	Suppressed	0.096
Clin oncology	Suppressed	0.032
ENT	0.1	0.128
Anticoagulant service	Suppressed	0.068

An entry was suppressed if it appeared in less than 5 patients.

A&E, accident and emergency; ENT, ear, nose, and throat; G Medicine, general medicine; ICD, International Statistical Classification of Diseases and Related Health Problems; iPAH, idiopathic pulmonary arterial hypertension.