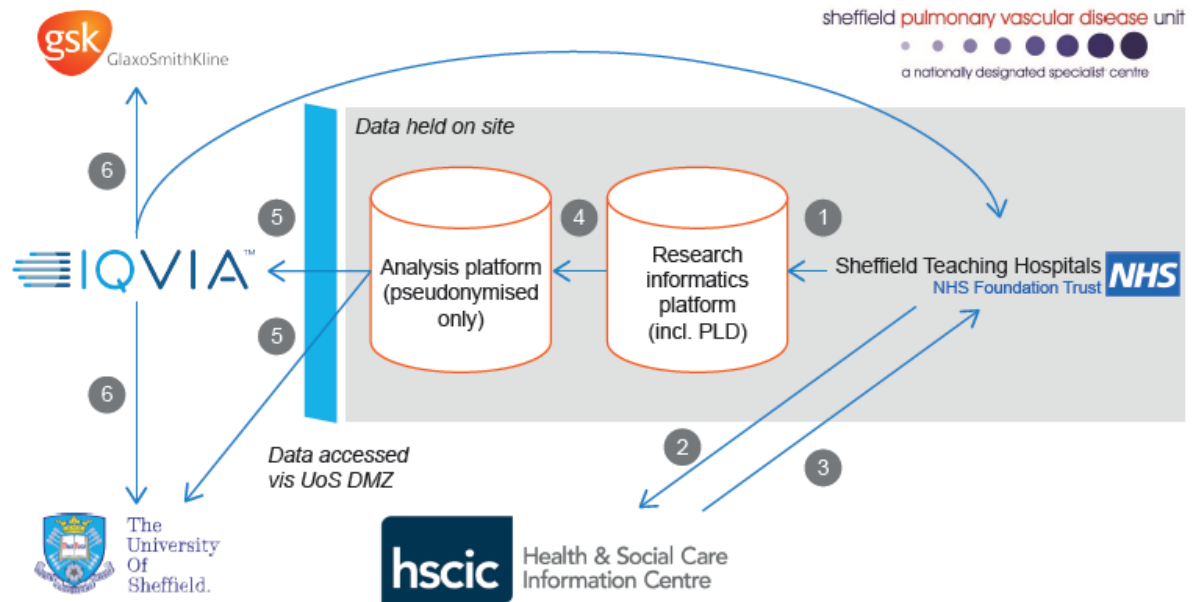


Supplementary Figure 1 Overview of data collection and ethical approvals at each stage of the analyses



Steps to Build the SPHInX Dataset

- 1 Sheffield teaching hospitals (STH) store clinical data on an internal research server at the Royal Hallamshire Hospital.
- 2 STH shares NHS numbers (matched to a study ID) for patients within the cohort with the Health & Social Care Information Centre (HSCIC).
- 3 HSCIC returns the requested de-identified HES records to STH research team with study ID and de-identified HES data for all non-STH patients.
- 4 Study ID linked HES data are handled in compliance with the STH policies on patient data, and are only accessible by the patient management team. STH research team de-identified the linked data which were deposited onto a server, physically secured within the STH Computer Sciences for analysis.
- 5 The data are accessed remotely within the STH DMZ by trained researchers (from IQVIA and specific members from the University of Sheffield (UoS) who are under honorary contracts at the STH). Access is granted using strong two-factor authentication based on USB keys which produce one time use passwords. Analysis was conducted for the agreed research questions and only on de-identified patient information.
- 6 IQVIA share aggregated longitudinal descriptive analysis of diagnostic pathways with GSK, STH and UoS

The SPHInX dataset was built under multiple different governance arrangements in order to ensure all research was undertaken on appropriate de-identified patient records. This included STH research ethics approval; National Health Research Authority research ethics and confidentiality advisory group approval, as well as approval of the Independent Group Advising on the Release of Data (via application to NHS Digital's Data Access Request Service)

Supplementary Table 1. ICD-10 diagnostic codes

ICD-10 Code	Description (<i>Grouping</i>)
<i>Due to left heart failure (increased back pressure in the pulmonary vessel)</i>	
I50.1	Left ventricular pump failure (heart attack, cardiomyopathy)
I11	Left ventricular stiffness (hypertension, diabetes, metabolic syndrome)
I35, I34	Valve disease (mitral or aortic stenosis or regurgitation)
<i>Diseases affecting the whole lung (lung diseases obliterate blood vessels)</i>	
J40, J41, J42, J43	Chronic bronchitis and emphysema (combination of loss of lung plus hypoxia)
J80, J81, J82, J84	Interstitial lung diseases (destructive diseases that obliterate vessels, such as pulmonary fibrosis, sarcoidosis, and many others)
<i>Hypoxia related (decreased oxygen constricts pulmonary blood vessels)</i>	
W94, T70	High altitude-dwelling
G47.3, R06.8	Sleep apnoea and other hypoventilation syndromes
Primarily obstructing diseases of the pulmonary vessels	
I26	Pulmonary thromboembolism

B65	Schistosomiasis
D57	Sickle cell anaemia
N/A	Tumour emboli
J98.5	Fibrosing mediastinitis (obstruction by fibrosis related to histoplasmosis)
D73	Haematological disorders: myeloproliferative disorders, splenectomy
D86, C96, D18.1, Q85.0, D69.0, L95, L94,	Systemic disorders: sarcoidosis, pulmonary Langerhans' cell histiocytosis, lymphangioleiomyomatosis, neurofibromatosis, vasculitis
E74.0, E75.2, E00-E07	Metabolic disorders: glycogen storage disease, Gaucher's disease, thyroid disorders
Y84.1	Others: Tumour obstruction, fibrosing mediastinitis, chronic kidney disease on dialysis
<i>Pulmonary arterial hypertension (changes in the structure and function of the pulmonary arteries)</i>	
N/A	Idiopathic (formerly primary pulmonary hypertension)
N/A	Heritable (formerly familial, due to BMPR2 or Alk-1 mutations)
N/A	Drug- and toxin-induced (stimulants)

M30, M31, M32, M33, M34, M35, M36	Connective tissue diseases (especially scleroderma)
B20, B21, B22, B23, B24	HIV infection (rare occurrence)
K76.6	Portal hypertension (cirrhosis and other advanced liver diseases)
Q20, Q21, Q22, Q23, Q24	Congenital heart disease that allows blood to shunt around the lungs
D18.0	Pulmonary veno-occlusive disease and pulmonary capillary haemangiomatosis (rare)

Alk-1, activin receptor-like kinase 1; BMPR2, bone morphogenetic protein receptor 2; HIV, human immunodeficiency virus; ICD-10, International Statistical Classification of Diseases and Related Health Problems, version 10

Supplementary Table 2. ICD-10 codes related to concomitant and differential diagnoses of PAH

ICD-10 code	Description
I42.0	Dilated cardiomyopathy
E03.9	Hypothyroidism
I05.0, I34.2 OR Q23.2	Mitral stenosis
M35.1	Mixed connective-tissue disease
G47.3	Obstructive sleep apnoea
M32	Systemic lupus erythematosus
K76.6	Portal hypertension
I37.0	Pulmonic stenosis
L94.0, L94.1 OR M43	Scleroderma
I20–I25	Ischaemic heart diseases
I50	Heart failure
I26–I28	Pulmonary heart disease and diseases of pulmonary circulation
J45	Asthma
J47 OR J40–J44	COPD
J84.9	Interstitial lung disease

COPD, chronic obstructive pulmonary disease; ICD-10, International Statistical Classification of Diseases and Related Health Problems, version 10; PAH, pulmonary arterial hypertension