Original research

# Topological data analysis identifies molecular phenotypes of idiopathic pulmonary fibrosis

Andrew Shapanis ,[1] Mark G Jones ,[2] James Schofield,[3] Paul Skipp [1]

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/thorax-2022-219731).

[1]Biological Sciences, University of Southampton, Southampton, UK
[2]Clinical and Experimental Sciences, Faculty of Medicine, University of Southampton, Southampton, UK
[3]TopMD Precision Medicine Ltd, Southampton, UK

**Correspondence to**
Prof Paul Skipp, Biological sciences, University of Southampton, Southampton, Hampshire, UK; pjss@soton.ac.uk

Received 10 October 2022
Accepted 19 January 2023

## ABSTRACT

**Background** Idiopathic pulmonary fibrosis (IPF) is a debilitating, progressive disease with a median survival time of 3–5 years. Diagnosis remains challenging and disease progression varies greatly, suggesting the possibility of distinct subphenotypes.

**Methods and results** We analysed publicly available peripheral blood mononuclear cell expression datasets for 219 IPF, 411 asthma, 362 tuberculosis, 151 healthy, 92 HIV and 83 other disease samples, totalling 1318 patients. We integrated the datasets and split them into train (n=871) and test (n=477) cohorts to investigate the utility of a machine learning model (support vector machine) for predicting IPF. A panel of 44 genes predicted IPF in a background of healthy, tuberculosis, HIV and asthma with an area under the curve of 0.9464, corresponding to a sensitivity of 0.865 and a specificity of 0.89. We then applied topological data analysis to investigate the possibility of subphenotypes within IPF. We identified five molecular subphenotypes of IPF, one of which corresponded to a phenotype enriched for death/transplant. The subphenotypes were molecularly characterised using bioinformatic and pathway analysis tools identifying distinct subphenotype features including one which suggests an extrapulmonary or systemic fibrotic disease.

**Conclusions** Integration of multiple datasets, from the same tissue, enabled the development of a model to accurately predict IPF using a panel of 44 genes. Furthermore, topological data analysis identified distinct subphenotypes of patients with IPF which were defined by differences in molecular pathobiology and clinical characteristics.

### WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Diagnosis of idiopathic pulmonary fibrosis (IPF) is challenging as there is currently no in vitro diagnostic test available. The clinical course of IPF is also highly heterogenous, suggesting the possibility of distinct subphenotypes.

### WHAT THIS STUDY ADDS

⇒ This study combines multiple publicly available peripheral blood mononuclear cell datasets of IPF and other diseases to create a prediction model which could accurately predict IPF in a diseased background to a high degree using a panel of 44 genes. Furthermore, topological data analysis revealed five distinct molecular phenotypes of IPF, which we characterise here using bioinformatic analysis.

### HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ The prediction model outlined here could be investigated further to help develop an in vitro diagnostic test for IPF. Additionally, the newly identified subphenotypes with the interpretation given here can help future research understand the heterogenous disease progression of IPF, ultimately leading to new, more targeted therapies.

## INTRODUCTION

Idiopathic pulmonary fibrosis (IPF) is a chronic and progressive disease of unknown cause resulting in significant fibrosis of the lungs, leading to a worsening of dyspnoea, lung function, and ultimately death.[1] The incidence of IPF has increased over time with cases currently ranging between 2.8 and 18 cases per 100 000 in Europe and North America.[2 3] The median age of IPF diagnosis is 65, with a median survival time from diagnosis of 3–5 years.[4]

IPF lacks an in vitro diagnostic test and in suspected IPF cases, diagnosis is reliant on high-resolution CT imaging and clinical expertise to identify a usual interstitial pneumonia (UIP) pattern, characterised by bilateral reticulation and honeycombing, typically in the lower lobes. If a UIP pattern is established, a detailed history is taken to identify any known causes (eg, mould in chronic hypersensitivity pneumonitis or asbestosis in pneumoconiosis),[5] but if no cause can be identified, an IPF diagnosis can be made. The differential diagnosis between other UIP diseases and IPF is a crucial one as this may influence treatment options and decisions.[6] Additionally, with the advent of new drugs (eg, pirfenidone (Esbriet) and nintedanib (Ofev)), there is a critical need for biomarkers for early diagnosis and for identifying patient groups where treatments will be most effective.

The conceptual belief of IPF as a chronic immune disease from the pre 2000's has shifted to one which involves abnormal chronic wound healing in response to consistent micro-epithelial injury.[6] Nonetheless, clinical progression of IPF is known to be heterogenous with some undergoing rapid progression, leading to a poor prognosis and early mortality, to some exhibiting very few exacerbations and a better prognosis.[7 8] The reason for the difference in progression is likely multifaceted but may be attributed to different disease subphenotypes, which although all present as fibrosis of the

Check for updates

**To cite:** Shapanis A, Jones MG, Schofield J, *et al*. *Thorax* Epub ahead of print: [*please include* Day Month Year]. doi:10.1136/thorax-2022-219731

lung, have been hypothesised to consist of different molecular phenotypes resulting in a heterogenous progression.

Topological data analysis (TDA) has been successfully used to discover information in large, sparse and complex biomedical datasets, including disease subphenotyping studies.[9–14] Arising from topology in applied mathematics, TDA provides a framework for analysing large volumes of high-dimensional data. In this study, we have used TDA to investigate for evidence of distinct molecular phenotypes of IPF using publicly available gene expression data of peripheral blood mononuclear cells (PBMCs). Using datasets from IPF and other heterogenous diseases, our study demonstrates that the transcriptomic profile of IPF is distinct from other diseases and using TDA identifies evidence for subphenotypes of IPF suggestive of differing underlying molecular pathophysiologies.

## MATERIALS AND METHODS

### Data collection

NCBI GEO study datasets corresponding to IPF PBMC expression data were collected (GSE38958,[15] GSE28042[16] and GSE132607[17]). Two other datasets were further collected for their utility as disease comparators relating to inflammation (GSE69683[18]: asthma) and infection (GSE37250[19]: HIV/ tuberculosis (TB)) in addition to the healthy patients included as part of these datasets. Online supplemental table 1 outlines the clinical characteristics of the IPF datasets.

### Normalisation and batch correction

Each dataset was log2 transformed (if not already done so) and digital precision normalised as described by Heider and Alt.[20] Genes with multiple probes in each dataset were averaged using the mean and all dataset expression matrices were combined into one large matrix. Genes which were present in <25% of the samples and samples containing <25% data were removed before finally removing all genes which had missing data to produce a full, complete expression matrix with no missing values. Data were then batch corrected between datasets using the ComBat function from the sva R package (V.3.44.0).

### Topological data analysis

TDA was performed on IPF and healthy expression data (across all datasets) using Ayasdi software (Ayasdi, Menlo Park, California, USA). A variance normalised Euclidean metric with L-Infinity centrality and Gaussian density lenses with resolution of 28 bins and gain of ×4, equalised was employed. Clusters across the TDA network were manually defined by selecting nodes which grouped together by edges per node (number of interconnected nodes).

### Differential expression, lung decline, pathway and CIBERSORT analysis

The R package, Limma (V.3.52.2), was used to identify differentially expressed genes (DEGs) between each IPF subphenotype and to all healthy patients. DEGs along with their Benjamini Hochberg adjusted p values and fold changes were analysed using ingenuity pathway analysis (IPA). Results were exported and combined in R. Lung decline was analysed using longitudinal data available in GSE132607 and plotted using Prism GraphPad. Forced vital capacity (FVC) and diffusing capacity for carbon monoxide (DLCO) values are given as predicted percentages. Pairwise comparisons between timepoints were performed with Tukey's correction. Longitudinal data were analysed by fitting a mixed model as implemented in GraphPad Prism. The mixed model uses a compound symmetry covariance matrix and is fit using restricted maximum likelihood. The p value for the null hypothesis was reported. Gene expression data for each sample was uploaded to CIBERSORT and analysed against the published LM22 immune signature allowing for 1000 permutations. Results were exported and grouped into the subphenotypes in R for plotting.

### Machine learning

Data were partitioned into training (67%) and test (33%) cohorts, maintaining the ratio of test:control (IPF:other) samples in each cohort. Online supplemental table 2 outlines the clinical characteristics of the train/test split. R packages, caret (V.6.0-92), pROC (V.1.18.0) and doparallel (V.1.0.17), Boruta (V 7.0.0), were utilised for machine learning. Boruta search and recursive feature elimination (RFE) was used to identify the 44 predictive genes in the training cohort. Boruta search was performed for 1000 iterations and genes scored for their importance allowing tentative genes to be carried forward. RFE was performed with 5× repeated, 10-fold cross-validation using random forest, for subsets of 10–50 genes. The subset of genes which provided the best area under the curve (AUC) in the training data were carried forward. The Boruta and RFE predictive genes were then combined and formed the 44 predictive gene panel which was used to train the support vector machine (SVM; with radial kernel) employing fivefold cross-validation repeated three times before testing it on the test cohort to see how well it predicts IPF/ no-IPF. Training was carried out measuring receiver operating characteristic curve AUC.

## RESULTS

We analysed publicly available PBMC expression datasets for 219 IPF, 411 asthma, 362 TB, 151 healthy, 92 HIV and 83 other disease samples, totalling 1318 patients. First, we investigated if PBMC derived gene expression data alone may be sufficient to accurately diagnose IPF. All datasets were combined and split randomly into training (67%) and test (33%) cohorts. RFE and Boruta search was performed on the training data to identify a panel of 44 predictive genes which were used to train a SVM with a radial basis function kernel. The resulting predictive model had an AUC of 0.964 (95% CI: 0.9452 to 0.9834) with a specificity and sensitivity of 0.89 and 0.865, respectively (figure 1A). Furthermore, when filtering the test set for only

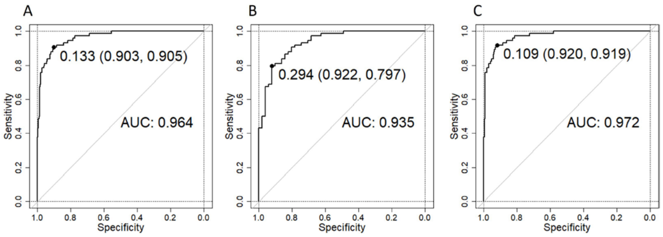


**Figure 1** Feature selection identified a panel of 44 genes capable of accurately predicting IPF. Data were split into training (67%) and test (33%) and automated feature selection used on the test set to identify predictive genes. Using these genes, a support vector machine with radial kernel was trained using the training set and tested on (A) complete test dataset; (B) only IPF and healthy patients in the test set and (C) only IPF and other diseases in the test set. Large black dots represent the optimal cut-off as per the maximum AUC (first value) followed by specificity and sensitivity in parentheses. AUC, area under the curve; IPF, idiopathic pulmonary fibrosis.

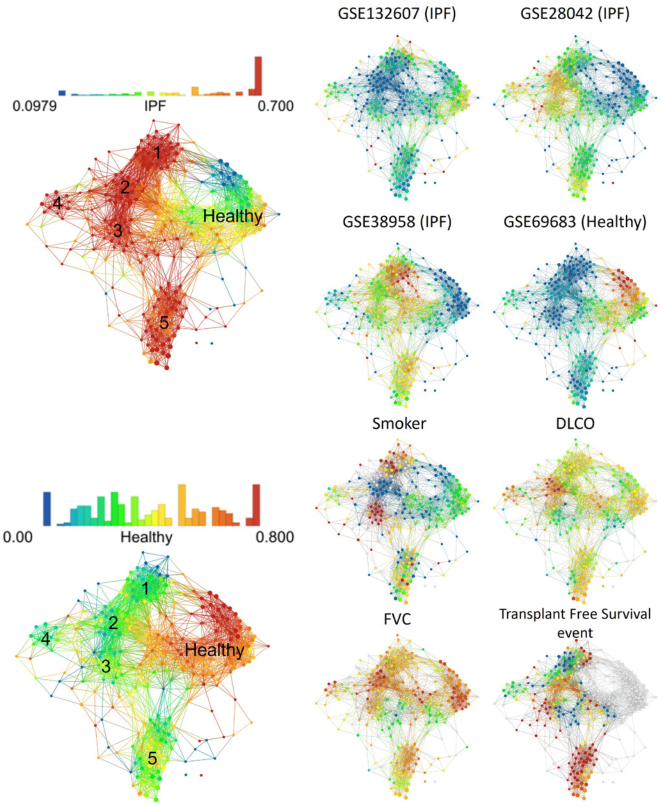


**Figure 2** TDA separates IPF from healthy individuals and highlights the presence of subphenotypes. IPF and healthy data from GSE132607, GSE28042, GSE38958 and healthy data from GSE69683 were subjected to TDA using Ayasdi software. A variance normalised Euclidean metric with L-infinity centrality and Gaussian distribution lenses, both with a resolution of 25 and a gain of 4 were used to generate the TDA structure. Dataset, smoker status, DLCO, FVC and transplant-free survival events were colour mapped onto the structure. red colouring denotes yes/high values; blue denotes no/low values. DLCO, diffusing capacity for carbon monoxide; FVC, forced vital capacity; IPF, idiopathic pulmonary fibrosis; TDA, topological data analysis.

IPF and healthy samples, an AUC of 0.935 (95% CI: 0.8937 to 0.976) with a specificity of 0.745 and sensitivity 0.892 was obtained (figure 1B). Moreover, when filtering the test set for only IPF and other diseases, an AUC of 0.972 (95% CI: 0.9549 to 0.9894) with specificity 0.939 and sensitivity of 0.865 was achieved (figure 1C). Thus, these findings suggest that the PBMC profile of patients with IPF is distinct. To discern uncertainty around the AUC, we also performed 10-fold cross-validation which produced an AUC of 0.954 (95% CI: 0.9280 to 0.9802) for all data, or an AUC of 0.930 (95% CI: 0.8885 to 0.9711) on only IPF and healthy or an AUC of 0.961 (95% CI: 0.9363 to 0.9863) on only IPF and other diseases.

## Topological data analysis

Having identified evidence of a distinct PBMC transcriptomic profile in patients with IPF, we then performed TDA across all gene expression data from IPF and healthy samples to investigate for the potential of subphenotypes of IPF. TDA identified five IPF subphenotypes with an even distribution of samples from each IPF dataset across the network, while in the 'healthy' cluster there was an enrichment of samples from dataset GSE69683, which was to be expected since only healthy patients were used from this dataset (figure 2 and online supplemental figure 1). Clinical features within the metadata included smoking history, DLCO, FVC and transplant-free survival can be found in online supplemental table 1. Mapping of these features onto the TDA network identified a cluster (subphenotype 5) enriched for death/transplant, with a significantly reduced time to event when compared with the other subphenotypes (online supplemental figure 2).

## GAP score and lung function

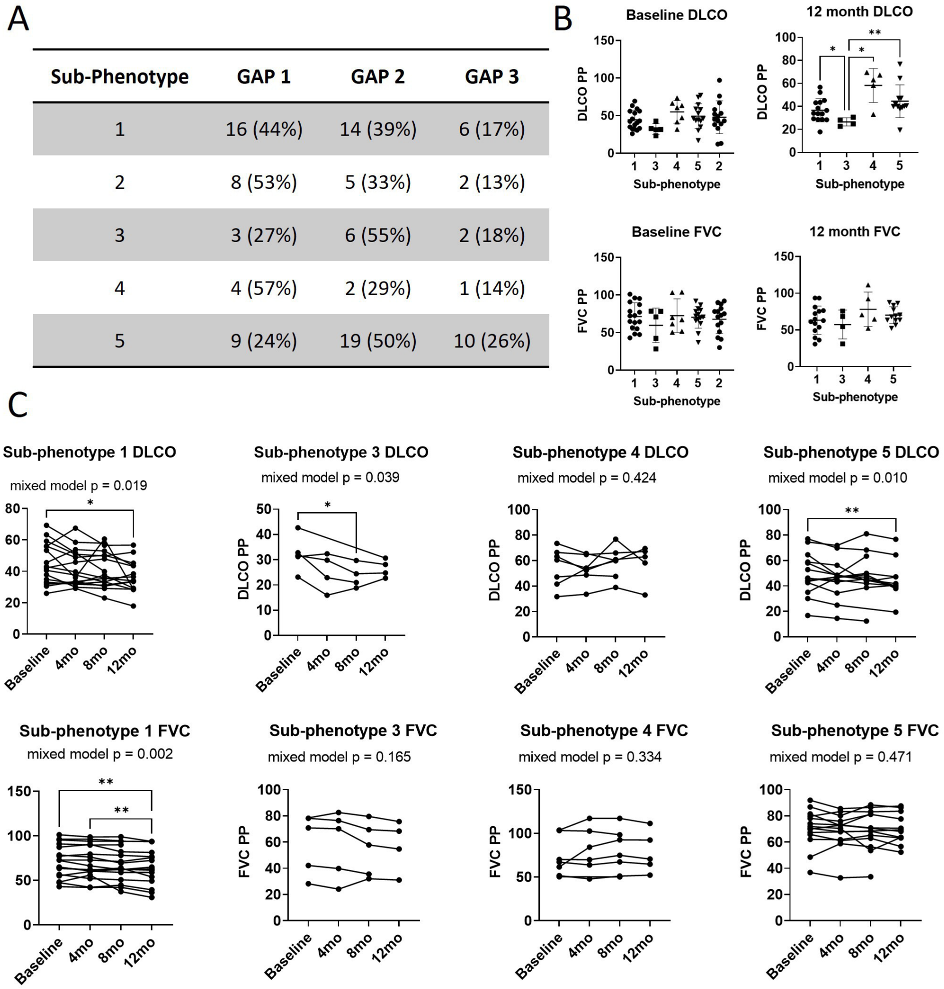
The Gender, Age and Physiology (GAP) index[21] for each patient was calculated using available data and split into their respective subphenotypes. No single subphenotype was heavily enriched for a particular GAP score, suggesting that the identified subphenotypes do not simply represent physiological or sex differences, although the IPF subphenotype enriched for death or transplant events (subphenotype 5), had a higher proportion of GAP stage 3 individuals while subphenotype 4 had the highest proportion of GAP stage 1 individuals (figure 3A).

No significant difference was observed between baseline FVC and DLCO data for each subphenotype (Brown-Forsythe and Welch analysis of variance tests, p values=FVC: 0.7895, DLCO: 0.1349), while at 12 months there was a significant reduction in DLCO for subphenotype 3 (figure 3B). To determine longitudinal lung function, data from GSE132607 was plotted (figure 3C), excluding subphenotype 2 where there were insufficient numbers. Subphenotype 1 had a significant decline in both DLCO and FVC between baseline and 12 months, while subphenotype 5 had a decline in DLCO between baseline and 12 months.

## Pathway and upstream regulator analysis

Significantly DEGs between each IPF subphenotype and all other patients with IPF, and between each IPF subphenotype and all other healthy patients were identified using the linear models for microarray and RNA-seq data (Limma) R package. Analysis of these DEGs using ingenuity pathway highlighted several key pathways and upstream regulators (figure 4). Comparative analysis of DEGs for each IPF subphenotype versus all healthy individuals highlighted several upstream transcriptional regulators that were common to all IPF subphenotypes (figure 4A). For example, each subphenotype showed a predicted activation of Hepatic Nuclear Factor 4 Alpha (*HNF4a*), dexamethasone, lipopolysaccharide, *TP53*, Nuclear Protein 1, Transcriptional Regulator (*NUPR1*), tretinoin or Spi-1 Proto-Oncogene (*SPI1*) and filgrastim, a drug used for the treatment of neutropenia. Predicted inhibition of cluster of differentiation 3 (*CD3*), MYC proto-oncogene, bHLH transcription factor (*MYC*), T-Cell receptor (*TCR*) and transcription factor 3 (*TCF3*) were seen within all subphenotypes. In contrast many of the upstream regulators, including *IL2*, *IL4*, Kruppel Like Factor 3 (*KLF3*), IL15, Transforming Growth Factor Beta 1 (*TGFB1*) and oestrogen receptor 1 (ESR1) showed differences in activation states (z-scores) across the subphenotypes suggesting unique mechanisms underpining the molecular pathobiology of the different subphenotypes.

Investigating the signalling pathways associated with the DEGs also revealed several pathways consistently downregulated across all subphenotypes (figure 4A), including PI3K signalling in B cells, telomerase signalling, 3-phosphoinositide biosynthesis and aldosterone signalling in epithelial cells. Many pathways were also differentially enriched between subphenotypes, including IL8, mTOR, TREM1, p53, HIF1a, B cell receptor and autophagy signalling. These differences in pathway activity



| Sub-Phenotype | GAP 1 | GAP 2 | GAP 3 |
|---|---|---|---|
| 1 | 16 (44%) | 14 (39%) | 6 (17%) |
| 2 | 8 (53%) | 5 (33%) | 2 (13%) |
| 3 | 3 (27%) | 6 (55%) | 2 (18%) |
| 4 | 4 (57%) | 2 (29%) | 1 (14%) |
| 5 | 9 (24%) | 19 (50%) | 10 (26%) |

**Figure 3** Each subphenotype is independent of GAP score and has its own lung function associated with it. (A) Using the available data, GAP scores were calculated for each subphenotype which emphasised that subphenotypes are novel and not dependant on GAP scores. (B) Baseline DLCO and FVC PP were plotted for each subphenotype at base line and 12 months. This showed no significant difference between any subphenotype except for subphenotype 3 which had a lower DLCO at 12 months. (C) Longitudinal data from GSE132607 was plotted for each subphenotype. Pairwise comparisons between each timepoint was performed with a Tukey's test. A mixed model was performed on each subphenotype and p value reported for each graph. Subphenotype 2 did not have enough patients from the GSE132607 dataset to investigate. DLCO, diffusing capacity for carbon monoxide; FVC, forced vital capacity; GAP, Gender, Age and Physiology; PP, predicted percentage.

between subphenotypes further suggests unique pathophysiobiology underpinning each of the subphenotypes (figure 4B).

We investigated expression of key genes identified in the publications associated with the integrated datasets used in this study[22 23] (figure 5). Subphenotype 1 showed higher expression of Inducible T-cell costimulator (*ICOS*), Tyrosine-protein kinase (*ITK*), lymphocyte-specific protein tyrosine kinase (*LCK*), Lysocardiolipin Acyltransferase 1 (*LCLAT1*), platelet-derived growth factor D (*PDGFD*) and β-galactosidase (*GLB1*), but lower expression of cyclin-dependent kinase inhibitor 1 (CDKN1A) and mucin 1 (MUC1). Subphenotypes 2, 3 and 4 showed similar high expression of *LCLAT1*, *GLB1*, and less *ICOS, ITK, LCK*,*CDKN1A and MUC1*, while subphenotype 5 had lower amounts of all except *CDKN1A*.

## CIBERSORT analysis

We next investigated for evidence of differences in immune cell composition between the identified subphenotypes. CIBERSORT analysis of gene expression values versus the LM22 CIBERSORT defined signature[24] were used to predict the immune cell proportions for each patient and then grouped into their respective subphenotypes (figure 6 and online supplemental table 3). Each subphenotype exhibited a specific immune cell profile with subphenotype 5 (that with highest incidence of death/transplant) having the highest proportion of activated mast cells and lowest proportion of naive CD4 T cells and activated dendritic cells. Subphenotype 1, that with the worst lung decline over time, had the highest number of naive CD4 T cells. These data suggest that each subphenotype is represented by a distinct immune cell profile.

## DISCUSSION

Here we integrated several PBMC transcriptomic datasets and identified evidence that patients with IPF have a distinct transcriptomic signature, with predictive modelling identifying a 44 gene signature capable of predicting IPF in a healthy and diseased population. Moreover, TDA revealed five novel subphenotypes of IPF. Importantly, this was all achieved within the same tissue type, PBMCs, where other studies that attempted this with less advanced clustering methods (hierarchical clustering, principle component analysis (PCA) and multidimensional scaling) and across inconsistent tissue subtypes. Each of these subphenotypes

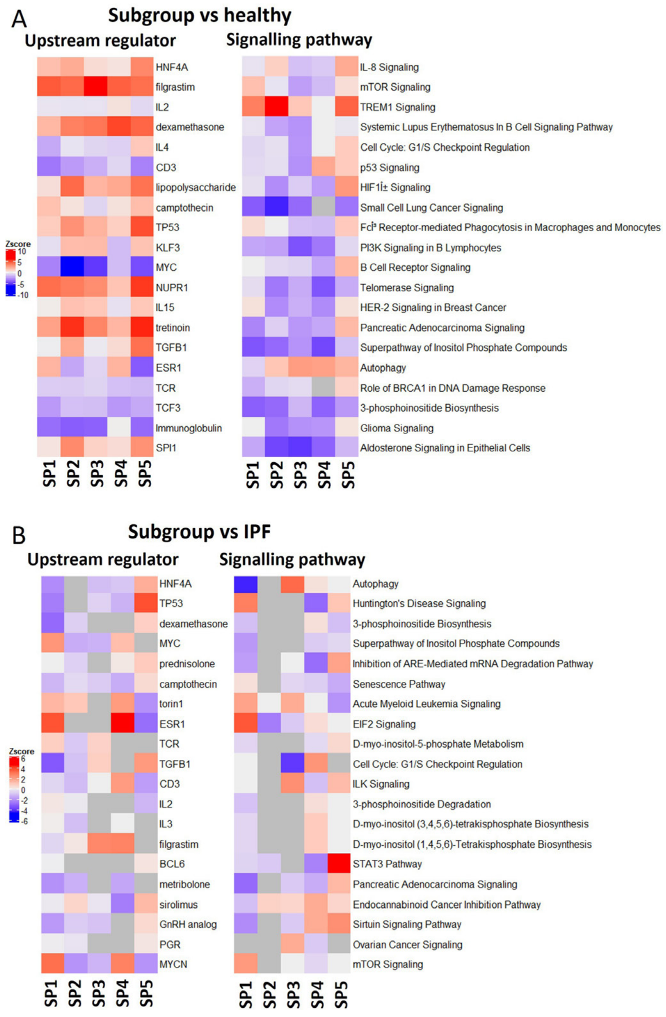


**Figure 4** Upstream regulator and pathway analysis reveals both commonalities and key differences between subphenotypes. Significantly differentially expressed genes for each subphenotype compared with(A), healthy patients or (B) all other patients with IPF. Significantly differentially expressed genes were analysed using ingenuity pathway analysis for upstream regulators (left heatmap of each panel) and pathway activity (right heatmap of each panel). Regulator and pathway analysis were combined into a heatmap and ordered by sum of log10 (p value) and filtered to top 20. IPF, idiopathic pulmonary fibrosis; SP, subphenotype.

were unique in their clinical and molecular characteristics, highlighting potential underlying mechanisms which could help future understanding and treatment of IPF.

Pathway and upstream regulator analysis highlighted key differences between the subphenotypes. HNF4A was strongly inhibited in subphenotype 1 and is an important transcription factor responsible for regulating several genes involved in liver development and may play a role in liver fibrosis. Induced expression of HNF4A in a Rat model of liver carcinogenesis has been reported to alleviate liver fibrosis[25 26] through conversion of myofibroblasts into hepatocyte-like cells.[27] Interestingly, HNF4A may also possess endocrine functions (or have endocrine effects) in systemic autoimmune rheumatic diseases[28] and as a major regulator of acute exacerbation of patients with chronic obstructive pulmonary disease (COPD).[29] Studies have suggested that 32%–35% of patients with IPF also have liver fibrosis[30] and that the overall survival was lower in those patients.[31] It is possible that HNF4A also contributes to lung fibrosis through a similar mechanism to liver fibrosis, modulating the immune system and myofibroblast dedifferentiation through common pathways.[32] PDGFD has been proposed as an indicator of liver fibrosis progression[33] and moreover when fibrosis was induced by bile duct ligation, PDGFD gene expression was shown to be significantly increased.[34] Interestingly, high systemic levels of PDGFD induced by hepatic adenovirus-based overexpression has been shown to be sufficient to induce interstitial kidney fibrosis and initiate fibroblast to myofibroblast differentiation,[35] which has also been shown to promote pulmonary fibrosis.[36] Here we identify that subphenotype 1 was also enriched for PDGFD expression, suggesting that this could represent a phenotype associated with multiorgan fibrosis.

Sex is a well-established risk factor of IPF with males having an increased likelihood of developing the disease.[21] Male mice with bleomycin induced pulmonary fibrosis present with more fibrosis than females. Interestingly, castrated male mice exhibited a response similar to female mice, while female mice given androgen exhibited a response similar to male mice, outlining the potential importance of sex hormones in exacerbating pulmonary fibrosis.[37] Premenopausal women have also been shown to have a lower risk of severe liver fibrosis than men, however, after menopause this risk becomes similar to the risk in men.[38] Our analysis predicted increased oestrogen receptor 1 (ESR1) activation for subphenotypes 1 and 4 and a decrease in subphenotypes 2, 3 and 5 when compared with healthy individuals. Subphenotype 5, the subphenotype which was enriched for people who either died or underwent transplant, had the lowest predicted activation of ESR1 across the subphenotypes, further suggesting a role for oestrogen and oestrogen-related signalling in IPF.

Given that IPF is a chronic and progressive disease, early detection and intervention is critical for prolonging life. Early intervention with nintedanib and pirfenidone have shown to reduce the rate of lung function decline, highlighting the need for early detection.[39–41] Currently, there is no approved blood based in vitro diagnostic and the gold standard remains multidisciplinary team discussions of high-resolution CT scans and in necessary cases, histological analysis of lung biopsies.[42] Herazo-Maya *et al*[2222] reported a panel of four genes for predicting IPF progression: *CD28*, *ICOS*, *LCK* and *ITK* where their higher expression was associated with a better transplant-free survival. This same dataset (GEO28042) also formed part of our integrated dataset, although CD28 was removed during the process of combining and batch correction. In line with Herazo-Maya, we observed lower expression of *ICOS*, *LCK* and *ITK* in subphenotype 5, the transplant and death subphenotype, but contrastingly found higher expression of these three genes in the subphenotype with the worst lung decline over time (subphenotype 1). Additionally, the study which generated dataset GSE38958 identified Lysocardiolipin Acyltransferase (LCLAT1) as having a positive effect on survival.[23] Our integrated analysis supports their findings with lower expression of LCLAT1 in subphenotype 5. However, similar to ICOS, LCK and ITK, we observed higher expression of LCLAT1 in subphenotype 1, that which has the worst lung decline over time. The reason behind these differences is unclear but may reflect our stratification of these individuals into subphenotypes, which was not the case for their studies. The high molecular weight glycoprotein KL-6 (MUC1) has previously been proposed as a blood marker of IPF, where an increase in MUC1 over time was associated with a lower survival.[43] However, the study sample size was limited to 145 patients and patients were treated with prednisone which current international consensus

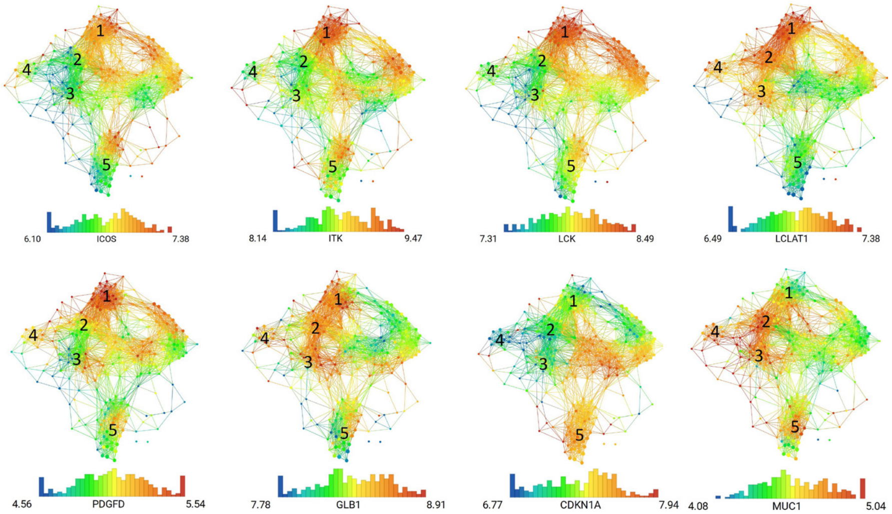


**Figure 5** Mapping expression of selected genes across the topological data analysis network highlights differences between subphenotypes.

guidelines do not endorse as a treatment for IPF. In our analysis, subphenotypes 2–4 have the highest MUC1 expression and so seemingly contradict these findings, however, given the lack of longitudinal data we are unable to investigate the change of MUC1 over the disease course.

A recent study investigated the proteome profile of blood plasma in 300 patients with IPF versus 100 healthy patients using the aptamer-based platform, SOMAscan.[44] The authors reported multiple prediction models for IPF diagnosis with 6 of 8 giving an AUC of 1. Although this study is promising and highlights the possibility of blood-based diagnostics of IPF, it suffers from several limitations. With aptamer-based technologies being expensive, the study's healthy cohort consisted of only 100 individuals and moreover, specifically excluded any control patients with respiratory diseases, cancer, autoimmune diseases, smokers and secondhand smoke sufferers. Although helpful in identifying mechanisms underpinning IPF, such a homogenous sample of the healthy population in a prediction model limits its likely ability to be translational and applicable to the wider population. Our study countered this issue through integration of multiple patient datasets including those with IPF, asthma, TB, HIV and several other diseases. One of the datasets used in our study identified a 52-gene panel that was able to predict poor outcome in IPF, specifically in the form of those that underwent transplant or died.[22] Given the absence of any in vitro diagnostic test for IPF, we decided to develop a general IPF/no-IPF predictive model to help towards solving this problem. Given that the panel outlined by Herazo-Maya *et al* was based on the prediction of poor outcomes in IPF, it is not too surprising that there are no shared genes between their 52-gene and our 44-gene panel.

Kraven *et al*[45] have recently taken a similar approach to our study, although there are several limitations that we believe are solved by our present study. Kraven *et al* combine several datasets to form a larger IPF dataset, but do so combining different tissue types. As the authors acknowledge, the integration of whole blood and PMBC which are significantly different may introduce

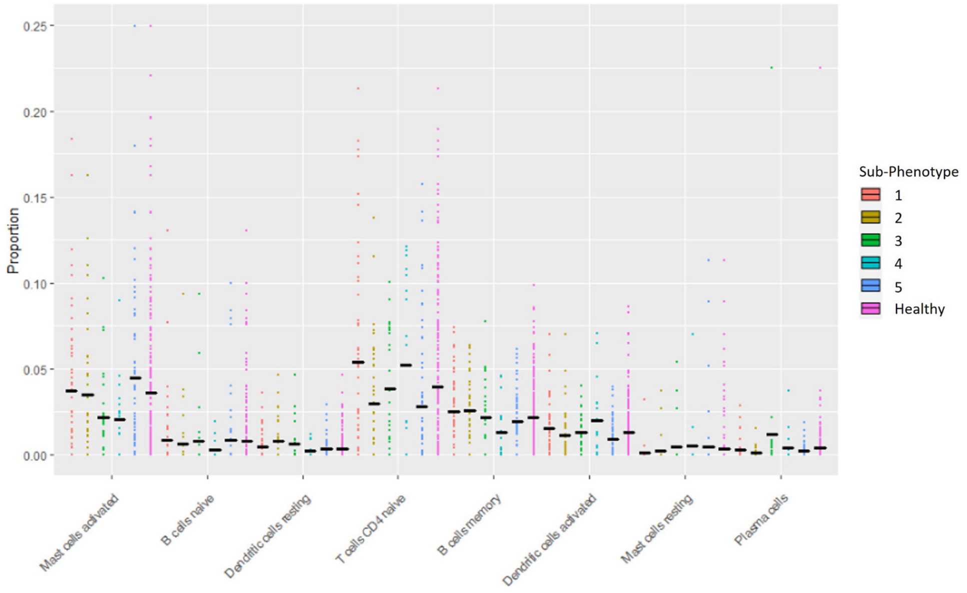


**Figure 6** CIBERSORT analysis reveals distinct immune signatures between subphenotypes. Gene expression data for each patient was analysed using CIBERSORT against the published LM22 immune signature, allowing 1000 permutations. Results were exported to R and plotted using ggplot2. Only signatures with significant difference are shown.

a systematic bias in the results. By comparison, we purposefully used data from one tissue type to reduce this potential bias. In our study, we used the unsupervised approach TDA, which has allowed the separation of patient data into subphenotypes without any predefined assumptions on the number of groups/clusters. This is in contrast to Karven *et al* who use a supervised clustering approach to predefine the number of clusters present and subsequently use PCA to assign each cluster. Clustering by PCA leads to a loss of data since only the principal components are used in the analysis, masking any non-linear relationships. We also biologically characterise these sub-phenotypes using a range of bioinformatic tools.

The findings of this study have to be seen in light of some potential limitations. Most notably not all clinical information was available for each of the datasets used, meaning that the time to event and treatment data were not available for each patient. Nonetheless, the original articles which produced the datasets do indicate that samples were collected treatment naive and so if present at all, will likely only be a few patients. We are also limited by the absence of a comparison to other interstitial lung diseases (ILDs), but unfortunately, there are currently no publicly available PBMC ILD datasets that we are aware of for comparison. Another limitation is that the predictive model outlined here has not been validated in a prospective cohort. We did however take a more conservative approach using both cross-validation and training/test split which informs us that the model is not overfitted to the data used for training. The limitations outlined here highlight the importance and need of a large prospective cohort of IPF and other ILDs.

In conclusion, we have integrated multiple datasets to develop a predictive model which can accurately predict IPF in a background of healthy patients, and patients with TB, HIV and asthma, using a panel of 44 genes, highlighting the potential for a non-invasive diagnostic tool for IPF. Furthermore, TDA identified distinct subphenotypes of patients with IPF which were defined by differences in molecular pathobiology and clinical characteristics.

**Contributors** AS collated and analysed the data. JS contributed and gave advice during the topological data analysis. MGJ provided clinical insight during the interpretation of the results. PS designed and supervised the study. All authors contributed to the drafting of the manuscript and provided critical feedback. AS and PS are guarantors of the paper.

**Funding** The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

**Map disclaimer** The inclusion of any map (including the depiction of any boundaries therein), or of any geographic or locational reference, does not imply the expression of any opinion whatsoever on the part of BMJ concerning the legal status of any country, territory, jurisdiction or area or of its authorities. Any such expression remains solely that of the relevant source and is not endorsed by BMJ. Maps are provided without any warranty of any kind, either express or implied.

**Competing interests** MGJ reports grants from Wellcome Trust, British Lung Foundation, Medical Research Council, Boehringher Ingelheim, AAIR and The Royal Society outside of the submitted work. JS reports consulting fees and shares in TopMD outside of the submitted work. PS reports consulting fees and shares in TopMD, outside of the submitted work.

**Patient consent for publication** Not applicable.

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Data availability statement** Data are available in a public, open access repository. All expression data in this study are freely available in Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/) with accession numbers: GSE38958, GSE28042, GSE132607, GSE69683, GSE37250.

**Supplemental material** This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

**ORCID iDs**
Andrew Shapanis http://orcid.org/0000-0003-4147-6956
Mark G Jones http://orcid.org/0000-0001-6308-6014
Paul Skipp http://orcid.org/0000-0002-2995-2959

## REFERENCES

1 Raghu G, Remy-Jardin M, Myers JL, *et al*. Diagnosis of idiopathic pulmonary fibrosis. An official ATS/ERS/JRS/ALAT clinical practice guideline. *Am J Respir Crit Care Med* 2018;198:e44–68.
2 Hutchinson J, Fogarty A, Hubbard R, *et al*. Global incidence and mortality of idiopathic pulmonary fibrosis: a systematic review. *Eur Respir J* 2015;46:795–806.
3 Hopkins RB, Burke N, Fell C, *et al*. Epidemiology and survival of idiopathic pulmonary fibrosis from national data in Canada. *Eur Respir J* 2016;48:187–95.
4 Ley B, Collard HR, King TE Jr. Clinical course and prediction of survival in idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med* 2011;183:431–40.
5 Lederer DJ, Martinez FJ. Idiopathic pulmonary fibrosis. *N Engl J Med* 2018;378:1811–23.
6 Selman M, King TE, Pardo A, *et al*. Idiopathic pulmonary fibrosis: prevailing and evolving hypotheses about its pathogenesis and implications for therapy. *Ann Intern Med* 2001;134:136–51.
7 Fernández Pérez ER, Daniels CE, Schroeder DR, *et al*. Incidence, prevalence, and clinical course of idiopathic pulmonary fibrosis: a population-based study. *Chest* 2010;137:129–37.
8 Martinez FJ, Safrin S, Weycker D, *et al*. The clinical course of patients with idiopathic pulmonary fibrosis. *Ann Intern Med* 2005;142:963–7.
9 Schofield JPR, Burg D, Nicholas B, *et al*. Stratification of asthma phenotypes by airway proteomic signatures. *J Allergy Clin Immunol* 2019;144:70–82.
10 Tariq K, Schofield JPR, Nicholas BL, *et al*. Sputum proteomic signature of gastro-oesophageal reflux in patients with severe asthma. *Respir Med* 2019;150:66–73.
11 De Meulder B, Lefaudeux D, Bansal AT, *et al*. A computational framework for complex disease stratification from multiple large-scale datasets. *BMC Syst Biol* 2018;12:60.
12 Bigler J, Boedigheimer M, Schofield JPR, *et al*. A severe asthma disease signature from gene expression profiling of peripheral blood from U-BIOPRED cohorts. *Am J Respir Crit Care Med* 2017;195:1311–20.
13 Östling J, van Geest M, Schofield JPR, *et al*. IL-17-high asthma with features of a psoriasis immunophenotype. *J Allergy Clin Immunol* 2019;144:1198–213.
14 Shapanis A, Lai C, Smith S, *et al*. Identification of proteins associated with development of metastasis from cutaneous squamous cell carcinomas (csccs) via proteomic analysis of primary csccs. *Br J Dermatol* 2021;184:709–21.
15 Zhou TZW, Ma SF, Wade M, *et al*. Profiling of gene expression in idiopathic pulmonary fibrosis. NCBI GEO. 2019. Available: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE38958
16 Herazo-Maya JD. Peripheral blood mononuclear cell gene expression profiles may predict poor outcome in idiopathic pulmonary fibrosis. NCBI GEO [agilent]. 2020. Available: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE28042
17 Yong H, Ma S, Martinez FJ, *et al*. Longitudinal blood transcriptomic changes predict lung function decline in idiopathic pulmonary fibrosis. NCBI GEO. 2022. Available: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE132607
18 Bigler J, Boedigheimer M, Schofield JPR, *et al*. Expression profiling in blood from subjects with severe asthma, moderate asthma, and non-asthmatics collected in the U-BIOPRED study. NCBI GEO. 2018. Available: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE69683
19 Anderson ST, Crampin AC, Dockrell HM, *et al*. Genome-wide transcriptional profiling of HIV positive and negative adults with active tuberculosis, latent TB infection and other diseases. NCBI GEO. 2020. Available: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE37250
20 Heider A, Alt R. VirtualArray: a R/bioconductor package to merge RAW data from different microarray platforms. *BMC Bioinformatics* 2013;14:75.
21 Ley B, Ryerson CJ, Vittinghoff E, *et al*. A multidimensional index and staging system for idiopathic pulmonary fibrosis. *Ann Intern Med* 2012;156:684–91.
22 Herazo-Maya JD, Noth I, Duncan SR, *et al*. Peripheral blood mononuclear cell gene expression profiles predict poor outcome in idiopathic pulmonary fibrosis. *Sci Transl Med* 2013;5:205ra136.

23 Huang LS, Mathew B, Li H, *et al*. The mitochondrial cardiolipin remodeling enzyme lysocardiolipin acyltransferase is a novel target in pulmonary fibrosis. *Am J Respir Crit Care Med* 2014;189:1402–15.

24 Newman AM, Liu CL, Green MR, *et al*. Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods* 2015;12:453–7.

25 Ning B-F, Ding J, Yin C, *et al*. Hepatocyte nuclear factor 4 alpha suppresses the development of hepatocellular carcinoma. *Cancer Res* 2010;70:7640–51.

26 Yue H-Y, Yin C, Hou J-L, *et al*. Hepatocyte nuclear factor 4alpha attenuates hepatic fibrosis in rats. *Gut* 2010;59:236–46.

27 Song G, Pacher M, Balakrishnan A, *et al*. Direct reprogramming of hepatic myofibroblasts into hepatocytes in vivo attenuates liver fibrosis. *Cell Stem Cell* 2016;18:797–808.

28 Hudson M, Bernatsky S, Colmegna I, *et al*. Novel insights into systemic autoimmune rheumatic diseases using shared molecular signatures and an integrative analysis. *Epigenetics* 2017;12:433–40.

29 Bertrams W, Griss K, Han M, *et al*. Transcriptional analysis identifies potential biomarkers and molecular regulators in pneumonia and COPD exacerbation. *Sci Rep* 2020;10:241.

30 Bocchino M, Brancaccio G, De Martino M, *et al*. Transient elastography detection of early liver fibrosis in idiopathic pulmonary fibrosis patients. *Eur Respir J* 2014;44:765.

31 Cocconcelli E, Tonelli R, Abbati G, *et al*. Subclinical liver fibrosis in patients with idiopathic pulmonary fibrosis. *Intern Emerg Med* 2021;16:349–57.

32 Makarev E, Izumchenko E, Aihara F, *et al*. Common pathway signature in lung and liver fibrosis. *Cell Cycle* 2016;15:1667–73.

33 Wang M, Gong Q, Zhang J, *et al*. Characterization of gene expression profiles in HBV-related liver fibrosis patients and identification of ITGBL1 as a key regulator of fibrogenesis. *Sci Rep* 2017;7:43446.

34 Borkham-Kamphorst E, van Roeyen CRC, Ostendorf T, *et al*. Pro-Fibrogenic potential of PDGF-D in liver fibrosis. *J Hepatol* 2007;46:1064–74.

35 Buhl EM, Djudjaj S, Babickova J, *et al*. The role of PDGF-D in healthy and fibrotic kidneys. *Kidney Int* 2016;89:848–61.

36 Li M, Luan F, Zhao Y, *et al*. Epithelial-Mesenchymal transition: an emerging target in tissue fibrosis. *Exp Biol Med (Maywood)* 2016;241:1–13.

37 Voltz JW, Card JW, Carey MA, *et al*. Male sex hormones exacerbate lung function impairment after bleomycin-induced pulmonary fibrosis. *Am J Respir Cell Mol Biol* 2008;39:45–52.

38 Yang JD, Abdelmalek MF, Pang H, *et al*. Gender and menopause impact severity of fibrosis among patients with nonalcoholic steatohepatitis. *Hepatology* 2014;59:1406–14.

39 King TE Jr, Bradford WZ, Castro-Bernardini S, *et al*. A phase 3 trial of pirfenidone in patients with idiopathic pulmonary fibrosis. *N Engl J Med* 2014;370:2083–92.

40 Noble PW, Albera C, Bradford WZ, *et al*. Pirfenidone in patients with idiopathic pulmonary fibrosis (capacity): two randomised trials. *Lancet* 2011;377:1760–9.

41 Richeldi L, du Bois RM, Raghu G, *et al*. Efficacy and safety of nintedanib in idiopathic pulmonary fibrosis. *N Engl J Med* 2014;370:2071–82.

42 Richeldi L, Collard HR, Jones MG. Idiopathic pulmonary fibrosis. *Lancet* 2017;389:1941–52.

43 Yokoyama A, Kohno N, Hamada H, *et al*. Circulating KL-6 predicts the outcome of rapidly progressive idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med* 1998;158:1680–4.

44 Todd JL, Neely ML, Overton R, *et al*. Peripheral blood proteomic profiling of idiopathic pulmonary fibrosis biomarkers in the multicentre IPF-PRO registry. *Respir Res* 2019;20:227.

45 Kraven LM, Taylor AR, Molyneaux PL, *et al*. Cluster analysis of transcriptomic datasets to identify endotypes of idiopathic pulmonary fibrosis. *Thorax* 2022. 10.1136/thoraxjnl-2021-218563 [Epub ahead of print 09 May 2022].