Forced vital capacity trajectories in patients with idiopathic pulmonary fibrosis: a secondary analysis of a multicentre, prospective, observational cohort


Summary

Background Idiopathic pulmonary fibrosis is a progressive fibrotic lung disease with a variable clinical trajectory. Decline in forced vital capacity (FVC) is the main indicator of progression; however, missingness prevents long-term analysis of patterns in lung function. We aimed to identify distinct clusters of lung function trajectory among patients with idiopathic pulmonary fibrosis using machine learning techniques.

Methods We did a secondary analysis of longitudinal data on FVC collected from a cohort of patients with idiopathic pulmonary fibrosis from the PROFILE study; a multicentre, prospective, observational cohort study. We evaluated the imputation performance of conventional and machine learning techniques to impute missing data and then analysed the fully imputed dataset by unsupervised clustering using self-organising maps. We compared anthropometric features, genomic associations, serum biomarkers, and clinical outcomes between clusters. We also performed a replication of the analysis on data from a cohort of patients with idiopathic pulmonary fibrosis from an independent dataset, obtained from the Chicago Consortium.

Findings 415 (71%) of 581 participants recruited into the PROFILE study were eligible for further analysis. An unsupervised machine learning algorithm had the lowest imputation error among tested methods, and self-organising maps identified four distinct clusters (1–4), which was confirmed by sensitivity analysis. Cluster 1 comprised 140 (34%) participants and was associated with a disease trajectory showing a linear decline in FVC over 3 years. Cluster 2 comprised 100 (24%) participants and was associated with a trajectory showing an initial improvement in FVC before subsequently decreasing. Cluster 3 comprised 113 (27%) participants and was associated with a trajectory showing an initial decline in FVC before subsequent stabilisation. Cluster 4 comprised 62 (15%) participants and was associated with a trajectory showing stable lung function. Median survival was shortest in cluster 1 (2.87 years [IQR 2.29–3.4]) and cluster 3 (2.23 years [1.75–3.84]), followed by cluster 2 (4.74 years [3.96–5.73]), and was longest in cluster 4 (5.56 years [5.18–6.62]). Baseline FEV1 to FVC ratio and concentrations of the biomarker SP-D were significantly higher in clusters 1 and 3. Similar lung function clusters with some shared anthropometric features were identified in the replication cohort.

Interpretation Using a data-driven unsupervised approach, we identified four clusters of lung function trajectory with distinct clinical and biochemical features. Enriching or stratifying longitudinal spirometric data into clusters might optimise evaluation of intervention efficacy during clinical trials and patient management.

Funding National Institute for Health and Care Research, Medical Research Council, and GlaxoSmithKline.

Copyright © 2022 The Author(s). Published by Elsevier Ltd. This is an Open Access article under the CC BY 4.0 license.
To mitigate this bias, previous studies have used various methods to adjust for data loss.\(^6,9\) However, these approaches can introduce alternative biases, making it difficult to accurately measure and model the disease trajectory of idiopathic pulmonary fibrosis over extended time periods.\(^6\)

Machine learning algorithms can overcome some assumptions and might mitigate biases induced by other imputation methods.\(^1\) Missing data remain an issue for machine learning tools; however, additional mathematical techniques can estimate numerous possible outcomes by resampling the underlying distributions thousands of times\(^1\) to generate enhanced synthetic datasets, which can be used to train machine learning algorithms\(^1\) to operate as an imputation tool.\(^2,19\)

We aimed to enhance the power of a longitudinal cohort of patients with incident idiopathic pulmonary fibrosis through imputation of data on lung function to estimate FVC loss due to disease progression. Subsequently, we applied unsupervised self-organising maps (SOMs) to identify distinct clusters of disease trajectories among patients with idiopathic pulmonary fibrosis, which could inform disease management and improve the efficacy of clinical trials.

**Methods**

**Study design**

We did a secondary analysis of longitudinal data on FVC collected from a cohort of patients with idiopathic pulmonary fibrosis from the PROFILE study,\(^19\) a multicentre, prospective, observational cohort study. We also performed a replication of the analysis on an independent dataset (ie, the replication cohort), obtained from the Chicago Consortium, which included longitudinal FVC measures obtained from the UUS study (in the USA, the UK, and Spain) collected by the University of Chicago (Chicago, IL, USA).\(^18\) The PROFILE study and the replication cohort have been described previously (appendix p 1).\(^19,20\)

**Data analysis**

Imputation methods were chosen on the basis of the data being continuous rather than categorical, and following literature review.\(^1,9,10,19-21\) Methods included simple
interpolation of missing values, including conventional linear regression, last observation carried forward, and 10% annual reduction in percentage predicted FVC (–10% decline per year), as well as machine learning approaches, including random forest and k-nearest neighbours classifiers capable of dealing with non-linear data and data that are not normally distributed. Due to the longitudinal connectivity between the spirometric visits related to a patient, all imputations were performed as consecutive chained equations.

For testing imputation methods, we used the complete dataset, consisting of 82 patients who completed all six spirometric visits (appendix p 13), split into learning datasets (57 [70%]) and test datasets (25 [30%]). Internal ten-fold cross-validation was used to optimise machine learning models. Synthetic simulation of missing data was conducted by removing data randomly from the test dataset, in proportion to the distribution of the occurrence of missing spirometric appointments in the whole PROFILE cohort. The lowest normalised root mean squared deviation (NRMSD) from separate models was used to assess the reliability of imputation. This index is used to measure the differences between predicted and observed values divided by the SD of the observed values.

To minimise survival bias and to increase statistical power, we did an analysis that included imputed values at all timepoints, regardless of the reason for missingness, including death. Based on the results of the complete dataset, we built a continuous autoregressive model. Integrating this model into Markov Chain Monte Carlo (MCMC) allowed incorporation of stochastic volatility over time, simulating events not experienced by patients in the complete dataset, such as abrupt FVC decreases or below-mean FVC values preceding patient death, which we termed the naive dataset. To mitigate against residual survival bias in this naive dataset, we generated a further theoretical dataset (10000 simulations each). In this dataset, we substituted 41.7% dummy values (including FVC=0) into the naive dataset and distributed these values proportionally to the mortality rate observed from the first year to the third year in the PROFILE study. We assessed the sensitivity of these imputation approaches by comparing NRMSD values across all spirometric visits.

We performed the unsupervised cluster analysis using SOMs. As a preprocessing step, we normalised the data by centralisation and scaling, which transformed the data into scale-free values. We performed hyperparameter optimisation before clustering. The SOM network was trained for the corresponding dataset for 200 iterations to minimise quantisation error. The learning rates started from 1.00 and was set to 0.90 (ordering) and to 0.02 (tuning), and a neighbourhood distance was set at 1.00 with hexagonal topology. Due to algorithmic similarities between k-means and SOMs, we used the Elbow method to identify the optimal number of clusters in our datasets. The validity (or stability) of each cluster was assessed by Jaccard indices after the sensitivity analysis. The minimum threshold for cluster stability by Jaccard indices was set at 50%.

We performed three additional sets of sensitivity analyses on the generated clusters. First, clusters were generated by use of 3 years of spirometry data from the following datasets: the complete PROFILE dataset, the complete PROFILE dataset excluding patients with data missing due to death, and data from patients who completed all spirometric visits without imputation. The second sensitivity test analysed the clusters generated by use of spirometry data from baseline to the first year, baseline to the second year, baseline to the third year, and from patients who completed all six spirometric visits. Theses analyses were performed in the same way in the replication cohort. The final sensitivity test included the cluster generation by k-means on the PROFILE dataset.

Serum biomarkers were measured from samples that were prospectively collected at baseline and analysed as previously described (appendix pp 1–2).

Statistical analysis
We implemented a workflow using open-source packages from the R project (version 4.1.1). Scripts are deposited online. To evaluate associations between lung function and disease trajectory between clusters, we applied a mixed-effects linear model with repeated measures analysis of annual rate of change in FVC. We performed the mortality risk assessment between clusters using hazard ratios (based on the Cox proportional hazards model), Kaplan-Meier plots, and log-rank tests. Survival probability at any particular timepoint was calculated by the formula: [(number of participants living at the start – number of participants who died) / number of participants living at the start].

Estimates for the Cox proportional hazards model and mixed-effects linear model tests were adjusted for covariance and limited to baseline percentile-predicted FVC in all analyses. Wilcoxon’s signed-rank test was used for continuous variables, and Fisher’s exact test was applied for categorical variables.

All comparisons among clusters were adjusted with the Bonferroni correction method. Data are median (95% CI), unless otherwise indicated. All statistical tests were two-sided, and p<0.05 was considered to be significant.

Role of the funding source
The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Results
415 (71%) of 581 participants recruited into the PROFILE study were eligible for this secondary analysis, while 180 (40%) of 455 participants from the independent
dataset were eligible for inclusion in the replication cohort (figure 1). Mean baseline FVC was 80·1% (SD 18·9). 321 (77%) participants were men, 94 (23%) were women, and mean age among participants was 70·6 years (SD 7·8; appendix p 13). Data on complete lung function were available in 82 (20%) participants. Data were missing due to death in 173 (42%) participants, of whom 48 (12%) died during the first year, 68 (16%) in the second year, and 57 (14%) in the third year. These missing data values meant that 488 (29·4%) of 1660 data points required imputation for the full analysis of lung function. A further 196 (11·8%) data points were missing for unknown reasons. Overall, from the complete PROFILE dataset of 415 patients, the dataset excluding patients with data missing due to death comprised 242 patients, and all data points were available from 82 patients who completed all spirometric visits.

Across all simulations, the random forest MCMC approach had the lowest mean NRMSD (0·4 [SD 0·2]), compared with models performing imputation by value interpolation and the other machine learning algorithm, k-nearest neighbours. Linear regression had the highest mean NRMSD (1·3 [0·8]) at each timepoint beyond 6 months, compared with all other methods (appendix p 3). In the sensitivity analysis that was performed for random forest MCMC imputation to compare unimputed data with naive and theoretical imputation models, there was little effect of imputation at 12 months (figure 2). After 2 years, differences were observed between the unimputed dataset (mean value 75·9 [95% CI 73·9–77·8]) and the naive model (69·5 [67·6–71·5]; p=0·0050), and between the unimputed dataset and the theoretical model (66·0 [63·9–68·0]; p<0·0001). The naive random forest imputation model had the lowest NRMSD mean value and was, therefore, used for further SOM cluster analysis (appendix p 4). The cluster sensitivity and validity analyses of these datasets suggested that the optimal number of discrete clusters derived using SOMs...
Cluster 1 was the largest cluster, with 140 (34%) of 415 participants, and was associated with a disease trajectory showing a linear decline in mean FVC over 3 years (table 1; figure 3A, B). Cluster 2 comprised 100 (24%) participants and was associated with a trajectory of improving lung function during the first year, before a subsequent decline in function in the second and third year (figure 3C, D). Cluster 3 comprised 113 (27%) participants and was associated with a trajectory of initial linear decline over the first year, before a subsequent decline in function over the second and third year (figure 3C, D). Cluster 3 comprised 62 (15%) participants and was associated with a trajectory showing a linear decline in mean FVC over 3 years (figure 3G, H).

In the replication cohort, comprising 180 individuals who qualified for imputation, the optimal number of clusters obtained was 0·75 (SD 0·20) for cluster 1, 0·64 (0·16) for cluster 2, 0·63 (0·10) for cluster 3, and 0·56 (0·10) for cluster 4 (appendix pp 5–6, 12).

Cluster 1 was the largest cluster, with 140 (34%) of 415 participants, and was associated with a disease trajectory showing a linear decline in mean FVC over 3 years (table 1; figure 3A, B). Cluster 2 comprised 100 (24%) participants and was associated with a trajectory of improving lung function during the first year, before a subsequent decline in function in the second and third year (figure 3C, D). Cluster 3 comprised 113 (27%) participants and was associated with a trajectory of initial linear decline over the first year, before a subsequent decline in function over the second and third year (figure 3C, D). Cluster 4 was the smallest cluster, comprising 62 (15%) participants, and was associated with a trajectory showing largely stable mean lung function over all 3 years (figure 3G, H).

In the replication cohort, comprising 180 individuals who qualified for imputation, the optimal number of clusters was also four (appendix p 8). SOM analysis showed similar cluster architecture with regard to the size of each cluster and the nature of lung function trajectories, with 74 (44%) participants in cluster 1, 38 (21%) in cluster 2, 42 (23%) in cluster 3, and 26 (14%) in cluster 4 (table 2; appendix p 8). Furthermore, the four clusters generated by SOMs in the PROFILE dataset were reproduced with the k-means clustering algorithm. These k-means clusters had identical architecture and similar membership allocation to those generated by SOMs (appendix p 12).

Participants in cluster 1 followed a linear decline in lung function, and this represented the most common phenotype in both the PROFILE cohort and the replication cohort (figure 4; table 1; appendix p 8). These patients had similar median survival with and without adjustment for baseline FVC in both cohorts (2·87 years [IQR 2·29–3·40]; figure 4; appendix p 10). In cluster 1, participants were generally younger and contained more never smokers, although the association between smoking status and disease trajectory was not significant (p=0·084). Biochemically, cluster 1 was associated with the highest concentrations of serum surfactant protein-D (SP-D; figure 5). Cluster 2 was the third most common cluster in both the PROFILE cohort and the replication cohort, and had a low number of never smokers in both cohorts (table 1). This cluster was associated with older age and a history of ever smoking. Concentrations of SP-D, as well as the FEV, to FVC ratio, were significantly lower in cluster 2 than in clusters 1 and 3 (figure 5A, D).

Table 1: Clinical characteristics per cluster in the PROFILE cohort

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>Cluster 1</th>
<th>Cluster 2</th>
<th>Cluster 3</th>
<th>Cluster 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of patients</td>
<td>415 (100%)</td>
<td>140 (34%)</td>
<td>100 (24%)</td>
<td>113 (27%)</td>
<td>62 (15%)</td>
</tr>
<tr>
<td>Age, years</td>
<td>70·6 (7·9)</td>
<td>69·7 (7·6)</td>
<td>71·5 (8·0)</td>
<td>71·7 (7·7)</td>
<td>69·0 (8·3)</td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>321</td>
<td>109 (78%)</td>
<td>82 (82%)</td>
<td>82 (73%)</td>
<td>48 (77%)</td>
</tr>
<tr>
<td>Female</td>
<td>94</td>
<td>31 (22%)</td>
<td>18 (18%)</td>
<td>31 (22%)</td>
<td>14 (23%)</td>
</tr>
<tr>
<td>Never smoker</td>
<td>117 (28%)</td>
<td>46 (33%)</td>
<td>24 (24%)</td>
<td>33 (29%)</td>
<td>14 (23%)</td>
</tr>
<tr>
<td>Ethnicity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>European descent</td>
<td>398 (96%)</td>
<td>134 (96%)</td>
<td>95 (95%)</td>
<td>110 (98%)</td>
<td>59 (95%)</td>
</tr>
<tr>
<td>Other</td>
<td>17 (4%)</td>
<td>6 (4%)</td>
<td>5 (5%)</td>
<td>3 (2%)</td>
<td>3 (5%)</td>
</tr>
<tr>
<td>Never used immunosuppressive medication</td>
<td>407 (98%)</td>
<td>137 (98%)</td>
<td>100 (100%)</td>
<td>111 (98%)</td>
<td>59 (95%)</td>
</tr>
<tr>
<td>Never used antifibrotic medication</td>
<td>303 (73%)</td>
<td>96 (70%)</td>
<td>70 (70%)</td>
<td>87 (77%)</td>
<td>50 (81%)</td>
</tr>
<tr>
<td>Completed all visits</td>
<td>82 (20%)</td>
<td>23 (16%)</td>
<td>24 (24%)</td>
<td>18 (16%)</td>
<td>17 (27%)</td>
</tr>
<tr>
<td>Missed visit at random</td>
<td>160 (39%)</td>
<td>44 (31%)</td>
<td>49 (49%)</td>
<td>32 (28%)</td>
<td>35 (56%)</td>
</tr>
<tr>
<td>Missed visit due to death</td>
<td>171 (42%)</td>
<td>73 (52%)</td>
<td>27 (27%)</td>
<td>63 (56%)</td>
<td>10 (16%)</td>
</tr>
<tr>
<td>Baseline percentile-predicted FVC</td>
<td>80·11 (19·23)</td>
<td>79·41 (19·80)</td>
<td>80·73 (18·69)</td>
<td>78·90 (18·92)</td>
<td>82·94 (19·49)</td>
</tr>
<tr>
<td>Baseline percentile-predicted diffusion capacity for carbon monoxide</td>
<td>46·03 (14·98)</td>
<td>43·24 (12·85)</td>
<td>48·66 (16·10)</td>
<td>44·42 (14·26)</td>
<td>51·31 (16·93)</td>
</tr>
<tr>
<td>Patients analysed due to missing data</td>
<td>380 (92%)</td>
<td>130 (93%)</td>
<td>89 (89%)</td>
<td>105 (93%)</td>
<td>56 (90%)</td>
</tr>
<tr>
<td>Baseline percentile-predicted FEV₁</td>
<td>82·70 (18·1)</td>
<td>83·57 (19·1)</td>
<td>83·4 (17·9)</td>
<td>81·1 (17·6)</td>
<td>82·6 (17·1)</td>
</tr>
<tr>
<td>Patients analysed due to missing data</td>
<td>413 (100%)</td>
<td>138 (99%)</td>
<td>100 (100%)</td>
<td>113 (100%)</td>
<td>62 (100%)</td>
</tr>
<tr>
<td>Frequency of MUC-5B allele</td>
<td>34·6 %</td>
<td>36·80 %</td>
<td>36·56 %</td>
<td>31·37 %</td>
<td>35·85 %</td>
</tr>
<tr>
<td>Patients analysed due to missing data</td>
<td>378 (91%)</td>
<td>125 (89%)</td>
<td>93 (93%)</td>
<td>102 (90%)</td>
<td>58 (94%)</td>
</tr>
<tr>
<td>&lt;5-year survival</td>
<td>149 (36%)</td>
<td>32 (23%)</td>
<td>47 (47%)</td>
<td>31 (27%)</td>
<td>39 (63%)</td>
</tr>
<tr>
<td>Survival, years</td>
<td>4·18 (3·87–4·37)</td>
<td>2·87 (2·29–3·40)</td>
<td>4·74 (3·96–5·73)</td>
<td>2·23 (1·75–3·84)</td>
<td>5·65 (5·18–6·62)</td>
</tr>
</tbody>
</table>

Data are n (%), mean (SD), or median (IQR). FVC=forced vital capacity. FEV₁=forced expiratory volume in 1 s.
Additionally, in the PROFILE cohort, participants in cluster 2 had significantly longer median survival (4.74 years [IQR 3.96–5.73]) than did those in cluster 1 (2.87 years [2.29–3.40]) and in cluster 3 (2.23 years [1.75–3.84]; p<0.0001; figure 4). The unadjusted median survival of participants in cluster 2 did not differ significantly from that of participants in cluster 1 or cluster 3 in the replication cohort (appendix p 9), but was similar to that of participants in cluster 2 in the PROFILE cohort when adjusted for baseline lung function (appendix p 10).

Participants in cluster 3 showed an initial decline in lung function with subsequent stabilisation, and this cluster was the second most common cluster in both cohorts (figure 3E, F; table 1). This cluster was associated with high mortality (figure 4; table 1), high FEV1, to FVC ratio (figure 5), and high concentrations of PRO-C28 (figure 5C). Similarly, high mortality was observed among participants in cluster 3 in the replication cohort (appendix p 10).

Cluster 4 represented the smallest group of patients in both cohorts and reflected stable lung function over 3 years.

Figure 3: SOMs of FVC trajectory per cluster among patients in the PROFILE cohort
Individual data points indicate the values obtained by the naive imputation model for each patient. Trendlines show mean (SD) ppFVC values at each timepoint. Individual spirometry traces clustered by SOMs from each patient are represented as scale-free normalised values for cluster 1 (A), cluster 2 (C), cluster 3 (E), and cluster 4 (G), and as non-normalised values for cluster 1 (B), cluster 2 (D), cluster 3 (F), and cluster 4 (H). Decline refers to percentage year decline in ppFVC from baseline. Significance tested following Bonferroni correction. SOM=self-organising map. FVC=forced vital capacity. pp=percentile-predicted. QE=quantisation error. NS=not significant.
Conventional linear regression was often limited to short durations. However, these studies did not seek to identify discrete patterns of disease trajectory. These clusters were associated with distinct anthropometric features with important implications for clinical management and future clinical trial design.

We applied a model-based cluster analysis that, following a series of internal sensitivity and validity analyses, showed four discrete clusters of lung function trajectory. These clusters were associated with distinct anthropometric features with important implications for clinical management and future clinical trial design.

Cluster analysis in interstitial lung disease is an emerging concept. At least three studies have performed such analyses using registry cohorts, integrating various clinical features (including comorbidities) in an attempt to identify distinct phenotypes.24-26 However, these studies did not seek to identify discrete patterns of disease trajectory. These clusters were associated with distinct anthropometric features with important implications for clinical management and future clinical trial design.

Cluster analysis in interstitial lung disease is an emerging concept. At least three studies have performed such analyses using registry cohorts, integrating various clinical features (including comorbidities) in an attempt to identify distinct phenotypes.24-26 However, these studies did not seek to identify discrete patterns of disease.

We applied a model-based cluster analysis that, following a series of internal sensitivity and validity analyses, showed four discrete clusters of lung function trajectory. These clusters were associated with distinct anthropometric features with important implications for clinical management and future clinical trial design.

Cluster analysis in interstitial lung disease is an emerging concept. At least three studies have performed such analyses using registry cohorts, integrating various clinical features (including comorbidities) in an attempt to identify distinct phenotypes.24-26 However, these studies did not seek to identify discrete patterns of disease.

We applied a model-based cluster analysis that, following a series of internal sensitivity and validity analyses, showed four discrete clusters of lung function trajectory. These clusters were associated with distinct anthropometric features with important implications for clinical management and future clinical trial design.

Cluster analysis in interstitial lung disease is an emerging concept. At least three studies have performed such analyses using registry cohorts, integrating various clinical features (including comorbidities) in an attempt to identify distinct phenotypes.24-26 However, these studies did not seek to identify discrete patterns of disease.

We applied a model-based cluster analysis that, following a series of internal sensitivity and validity analyses, showed four discrete clusters of lung function trajectory. These clusters were associated with distinct anthropometric features with important implications for clinical management and future clinical trial design.

Cluster analysis in interstitial lung disease is an emerging concept. At least three studies have performed such analyses using registry cohorts, integrating various clinical features (including comorbidities) in an attempt to identify distinct phenotypes.24-26 However, these studies did not seek to identify discrete patterns of disease.

We applied a model-based cluster analysis that, following a series of internal sensitivity and validity analyses, showed four discrete clusters of lung function trajectory. These clusters were associated with distinct anthropometric features with important implications for clinical management and future clinical trial design.

Cluster analysis in interstitial lung disease is an emerging concept. At least three studies have performed such analyses using registry cohorts, integrating various clinical features (including comorbidities) in an attempt to identify distinct phenotypes.24-26 However, these studies did not seek to identify discrete patterns of disease.

We applied a model-based cluster analysis that, following a series of internal sensitivity and validity analyses, showed four discrete clusters of lung function trajectory. These clusters were associated with distinct anthropometric features with important implications for clinical management and future clinical trial design.
behaviour in patients with idiopathic pulmonary fibrosis. Our analyses identified four distinct FVC trajectories, which challenge the current understanding of the natural history of idiopathic pulmonary fibrosis. Patients in clusters 1 and 3 showed disease trajectories that followed the expected decline in lung function over the first year, and this continued throughout the duration of illness for patients in cluster 1, but stabilised for patients in cluster 3. Patients in clusters 1 and 3 were, unsurprisingly, more likely to have data missing due to death. More surprisingly, a third of patients in the overall cohort followed an alternative trajectory (ie, clusters 2 and 4) and showed either improved or stable lung function in the first year followed by a conventional trajectory (cluster 2), or remained stable throughout the duration of the study (cluster 4). Clusters 2 and 4 were associated with a better prognosis in patients with incident idiopathic pulmonary fibrosis than were clusters 1 and 3. Similar findings were found in a post-hoc analysis of the INPULSIS studies, which investigated the efficacy and safety of nintedanib added to pirfenidone in patients with idiopathic pulmonary fibrosis. However, this analysis was performed without imputation, which might underestimate the effect in patients receiving placebo and lead to immortal time bias in favour of therapy, thus reinforcing the need to undertake imputation in such analyses. The reasons behind the improvement in lung function among participants in cluster 2 are unclear, but there are several possible explanations. These patients might have had acute, or infective, exacerbations at enrolment into both studies that improved before the typically observed deterioration in lung function occurred. Nevertheless, this potential reason is unlikely given that it would require over 20% of patients with idiopathic pulmonary fibrosis to have an acute or infective exacerbation within a 6-month period of enrolment into both studies. Although such exacerbations are common, most estimates of incidence of acute exacerbations are lower than 20% within 1 year and are associated with poor prognosis. Another explanation could be that cluster 2 included patients with concomitant chronic obstructive pulmonary disease who showed labile results on spirometry. Compared with the other three clusters, cluster 2 contained more ever smokers and the FEV1:FVC ratio was lower; however, this cluster was not associated with a lower diffusion capacity for carbon monoxide, the form of chronic obstructive pulmonary disease most commonly associated with idiopathic pulmonary fibrosis. The disease trajectory in cluster 2 might reflect response to antifibrotic or immunosuppressive therapy, although patients in both studies were not receiving antifibrotic therapy at the time of recruitment, and treatment in idiopathic pulmonary fibrosis.
fibrosis slows disease progression, rather than improves lung function. Furthermore, it is possible that individual variation in FVC values might have resulted in unusual patterns of lung function following cluster analysis; however, this is unlikely given the large number of patients in cluster 2 and that the size of the cluster were replicated in the replication cohort. Although the reasons for the observed increase in FVC over the first year in cluster 2 are yet to be elucidated, it is important to recognise its occurrence in a substantial proportion of patients with idiopathic pulmonary fibrosis. Failure to recognise this occurrence could mislead interpretation of clinical response if unequal randomisation occurs in trials. Importantly, in the CAPACITY1 study and the CAPACITY2 study, the groups receiving placebo showed a different disease trajectory to those receiving pirfenidone, which ultimately delayed the regulatory approvals and introduction of this treatment into clinical practice. It is possible that this finding might have been due to the inclusion of patients showing cluster 2 or 4 trajectories in the placebo groups, who, combined, made up 40% of the patient cohort with idiopathic pulmonary fibrosis in both the PROFILE cohort and the replication cohort.

Identifying patients who are likely to show the disease trajectories of clusters 2 and 4 could have practical implications for clinical management. The notion of a therapeutic trial would be misleading for patients in cluster 2, who are likely to show an improvement in FVC despite, rather than because of, therapy. Furthermore, the risk–benefit ratio of any given therapy might be altered among patients in cluster 4, particularly those recently diagnosed with an FVC of more than 80%. However, further studies to prospectively test the predictive power of such models are needed.

There are various strengths to the approach used in this study. We used several validity and sensitivity analyses to identify optimal imputation methods over both the short and long term. Importantly, when trying to define natural history, we were able to analyse data from a prospective cohort of largely untreated patients to generate the imputation models and to identify the clusters of lung function trajectory. Additionally, we were able to replicate the clusters’ architecture in an external cohort of patients with idiopathic pulmonary fibrosis in both the PROFILE cohort and the replication cohort.

This study identifies distinct trajectories of lung function in patients with idiopathic pulmonary fibrosis and has important implications for the development of clinical trials and clinical practice. Further improvement in collection of patient registry data and cluster methodology, as well as collaboration between research groups, will increase the accuracy of imputation and granularity of cluster analysis, thus facilitating further understanding of unique clusters of patients with pulmonary fibrosis, including those with pulmonary fibrosis of known cause. Development of these approaches could help to treat each patient with the correct treatment at the correct time.

Contributors

HPF, TMM, LVW, IDS, IT, and RGJ participated in the design and planning of the study. RGJ, GS, ER, RB, AUW, JP, IN, JMO, PLM, and TMM participated in recruitment of study patients and collected the data. WAF, MAK, RGJ, DJL, and EO developed and conducted the neoepitope and ELISA assays. HPF, IDS, IT, IN, ER, LMK, LVW, RJA, JMO, PLM, TMM, and RGJ did the data analysis and prepared the report for publication. HPF, IT, IDS, and RGJ were involved in the development of the study, implementation of the analytical workflow, machine learning optimisation, and final statistical analysis. HPF, IDS, and RGJ were involved in all stages of study development and delivery. All authors had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Declaration of interests

RGJ has received industry-academic funding from AstraZeneca, Biogen, Galecto, GlaxoSmithKline, Redx Pharma, and Pliant; and has received...
consultancy or speaker fees from Chiesi, Roche, PatientMPower, Bristol Myers Squibb, Veracyte, Daewoong, Resolution Therapeutics, RedX, Pliant, Boehringer Ingelheim, Galapagos, Vicore, and AstaZena. TMM has received industry-academic funding from GlaxoSmithKline and UCB; and has received consultancy or speaker fees from Apellis, AstraZeneca, Bayer, Biogen Idec, Boehringer Ingelheim, Celgene, Galapagos, GlaxoSmithKline, Indalo, Pliant, ProMetis, Roche, Samumed, and UCB. AWWW reports fees for work on scientific advisory boards for Intermune, Boehringer Ingelheim, Actelion, Gilead, Genentech, MedImmune, and Takeda; and lecturing fees from Intermune, Boehringer Ingelheim, and Actelion, outside of the submitted work. JMO reports an unrelated grant from Boehringer Ingelheim; and has received personal fees from Genentech, United Therapeutics, Gatehouse Bio, and AmMax Bio, unrelated to the submitted work. PLM received industry-academic funding from AstraZeneca via his institution; and has received speaker and consultancy fees from Boehringer Ingelheim and Hoffman-La Roche, outside of the submitted work. LVW received via his institution industry-academic funding from GlaxoSmithKline and holds British Lung Foundation Chair in Respiratory Research (CP–J). MAK is an employee and shareholder of Nordic Biosciences; and holds patents relating to neoptoepitopes. DJL is an employee and shareholder of Nordic Biosciences. JMBS is an employee and shareholder of Nordic Biosciences. WAF and ED are employees and shareholders of GlaxoSmithKline. All other authors declare no competing interests.

Data sharing

The PROFILE (anonymously processed) dataset used in this study can be made available on a reasonable request basis, which must include an appropriate protocol, analysis plan, and data exchange with institutional approvals in place before data transfer of any information. This request needs to be formally addressed to the head of the Margaret Turner Warwick for Centre for Fibrosing Lung Disease, RGC (g.s.jenkins@imperial.ac.uk). External validation data sources will not be provided because these are withheld by owners. Patient personal information and correspondence will not be provided because these are withheld by the corresponding author’s organisation to preserve patient privacy. The code for the method is freely available.

Acknowledgments

The PROFILE study was funded by the Medical Research Council (G09010226), National Institute for Health and Care Research (RP-2017­08­ST2–014), and GlaxoSmithKline (CRTJ14316), and was sponsored by Nottingahm University, Royal Brompton, and Harefield NHS Foundation Trust. HPF and RGF are funded by a National Institute for Health and Care Research (NHR) Research Professorship (RP-2017­08­ST2–014). Investigators received no financial incentives from the funding sources. JMO is supported by the National Institutes of Health–National Heart Lung and Blood Institute grant (R21HL138190, R01HL157895, and R36HL158933). PLM is supported by an Action for Pulmonary Fibrosis Mike Bray fellowship. RJA is funded by Action for Pulmonary Fibrosis as a Mike Bray Research Fellow. TMM is supported by an NIHR Clinician Scientist Fellowship (CS-2013­11­017). JP and RB are supported by the NIHR (National Institute for Health and Care Research) and University of Nottingham COVID­19 testing service. IT is supported by the University of Nottingham. LMK holds a studentship funded by the Medical Research Council. LVW is supported by the NIHR Leicester Biomedical Research Centre. IN is funded by a National Institutes of Health grant (NHHLBI HL145266). IDS holds an advanced Research Fellow (Rayne Foundation). It is partly supported by projects A­TIC­434­UCR20 (EU) and PID 2020­117978GB­100. The authors would like to thank L Knifton for her skilful wordsmithing and proofreading, GlaxoSmithKline and Nordic Biosciences provided invaluable support in the processing and quality control of clinical and biomarker data. The authors thank the patients, carers, families, health professionals, and everyone who participated in the PROFILE study.

References


30 Kwiatkowska S. IPF and CPFE - the two different entities or two different presentations of the same disease? *Adv Respir Med* 2018; 86: 23–26.