ERJ open research
THE BEST IN OPEN ACCESS BASIC, TRANSLATIONAL & CLINICAL RESPIRATORY RESEARCH

Early View

Original article

IPF cluster analysis highlights diagnostic delay and cardiovascular comorbidities association with outcome

J. Bordas-Martínez, R. Gavaldà, J.G. Shull, V. Vicens-Zygmunt, L. Planas-Cerezales, G. Bermudo-Peloche, S. Santos, N. Salord, C. Monasterio, M. Molina-Molina, G. Suarez-Cuartin

Please cite this article as: Bordas-Martínez J, Gavaldà R, Shull JG, *et al*. IPF cluster analysis highlights diagnostic delay and cardiovascular comorbidities association with outcome. *ERJ Open Res* 2021; in press (https://doi.org/10.1183/23120541.00897-2020).

IPF cluster analysis highlights diagnostic delay and cardiovascular comorbidities association with outcome

Take Home Message: Diagnostic delay and cardiovascular comorbidities impacts on IPF outcomes

+Bordas-Martínez J.[1,4], Gavaldà R.[2,3], Shull JG.[1], Vicens-Zygmunt V. [1], Planas-Cerezales L. [1], Bermudo-Peloche G.[1], Santos S. [1,4], Salord N. [4], Monasterio C. [4], * Molina-Molina M.[1], Suarez-Cuartin G.[1]

+ First autor

***Corresponding autor**

1. Interstitial Lung Disease Unit. Respiratory Department. Bellvitge University Hospital. IDIBELL. University of Barcelona. Hospitalet de Llobregat (Barcelona), Spain.
2. Amalfi Analytics, Barcelona, Spain
3. Computer Science department. Polytechnic University of Catalonia. Barcelona, Spain.
4. Sleep Unit. Respiratory Department. Bellvitge University Hospital. IDIBELL. University of Barcelona. Hospitalet de Llobregat (Barcelona), Spain.

Corresponding author;
Maria Molina-Molina, MD, PhD
Chief of the ILD Unit. Respiratory Dpt.
University Hospital of Bellvitge. IDIBELL
Hospitalet de Llobregat, 08907 (Barcelona)
mariamolinamolina@hotmail.com
tf. number: 0034 932607689

Key words: cluster analysis; idiopathic pulmonary fibrosis; diagnostic delay; cardiovascular comorbidity

Introduction

Idiopathic pulmonary fibrosis (IPF) is the most frequent and lethal interstitial lung disease (ILD) [1]. The development of antifibrotic treatments have increased the expected survival of IPF patients [2]. Given the overall poor quality of life of these patients, holistic care could impact daily life expectations. In order to improve current patient approaches, it is necessary not only to understand the disease, but also to assess different aspects of individual patients. The study of comorbidities [3–6], lifestyle, and psycho-emotional accompaniment for patients and family members [7] is fundamental in the comprehensive patient treatment. In this regard, different multivariable risk prediction models such as gender-age-physiology model (GAP) [8] or more recently the TORVAN model [5], have been created to predict the risk of death using different clinical data, lung functional tests and the presence of comorbidities. However, several other patient characteristics and health care features may have a significant role in disease outcome and patient needs.

Due to the heterogeneity of IPF presentation and progression, different phenotypes related to disease behavior and comorbidities have been explored [3, 4]. Some proposed phenotypes that present specific disease behavior are rapidly progressive IPF and combined pulmonary fibrosis and emphysema (CPFE) [3, 9]. On the other hand, the impact of comorbidities on disease behavior and mortality has been explored, proposing the term “comorbidome” [10]. Most IPF cases present with more than 2 comorbidities [10]. Cardiovascular diseases, pulmonary hypertension and lung cancer are the comorbidities with the highest impact on IPF mortality [10, 11]. However, some biological disorders may be a common trigger of different comorbidities in the same patient, such as metabolic syndrome or telomeric disorders [3, 9]. Cluster analysis has become a useful resource to identify homogeneous patients with similar clinical characteristics, prognosis and healthcare requirements [12, 13]. Additionally, the integrated study of respiratory diseases through clusters [14] helps identify hidden and unsuspected associations between different diseases and patient features, which could generate new hypotheses to be later explored in controlled studies. Previous analysis of chronic ILDs suggested distinct phenotypes which identified some meaningful clinical outcomes independent of disease diagnosis [13]. Furthermore, the better understanding of IPF patient profiles, including the different components that could influence patient needs such as disease behavior, comorbidities and patient condition, would optimize patient management.

Therefore, the aim of this study is to find hidden and/or unexpected associations in clusters of IPF patients based on common disease and patient features.

Methodology

This is an observational retrospective study analysing IPF patients treated with antifibrotic therapy. 136 IPF patients were treated with pirfenidone or nintedanib at the ILD Unit of Bellvitge Hospital (Barcelona, Spain) from 2012 to 2018. Of these, 6 were followed-up in another center. The diagnosis[1] and treatment[15] of the 130 included IPF patients were performed according to the ATS/ERS criteria[9] by the multidisciplinary committee.

Demographic variables collected were age, gender, body mass index (BMI), previous exposures (smoker habit, occupational and environmental exposures), clinical data (dyspnea, cough, crackles, nail clubbing), family history, comorbidities, pharmacological treatments, radiological pattern and hiatal hernia[16] on chest high-resolution computed tomography (HRCT) (hiatal hernia type II-IV with presence of air and food/fluid/air-fluid level in the esophagus were considered moderate and severe hiatal hernia respectively)[16], laboratory tests, sleep study (video-polysomnography or respiratory polygraphy), echocardiography, telomere length (TL) and lung biopsy when required. TL analysis was performed using DNA samples isolated from mouth epithelial cells

(oral swabs - Isohelix, SK-2S, Cell Projects Ltd) and peripheral blood mononuclear cells (Isohelix, Cell Projects Ltd)[17]. TL was considered shortened when z-score was below 25th percentile, and severe telomere shortening when below the 10th percentile [17]. Patients underwent pulmonary function tests (PFTs) including body plethysmography and spirometry, and 6 minutes walking tests (6MWT) at the time of diagnosis and thereafter every 3 months. Furthermore, forced vital capacity (FVC) and diffusing capacity for carbon monoxide (DLCO) were collected before starting antifibrotic treatment. Frequent respiratory infections were defined when > 2 respiratory infections with antibiotic requirement per year were present. Acute exacerbations that required hospital admission were defined following the current recommendations regardless of the trigger[18]. Antifibrotic treatment (pirfenidone and nintedanib), adverse events and subsequent management were followed for one year. Family aggregation, comorbidities, treatment-related side effects and drug compliance, and lung transplant or death due to IPF were recorded. Disease progression was defined as FVC decline ≥ 10% of predicted or DLCO ≥ 15% of predicted in 1 year. Progression-free survival (PFS) after 3-year follow-up was defined as no progression, lung transplant or death in 3 years of follow-up.

This study was approved by the Ethics Committee of Bellvitge University Hospital (reference code PR413/18). The study was performed in accordance with the ethical principles of the Declaration of Helsinki, and local laws for countries in which the research was done. Informed consent was obtained from each participant by the study investigator before patient data collection was done.

Cluster and statistical analysis

Clustering was performed using the MATE tool by Amalfi Analytics. Patients were clustered using approximate singular value-based tensor decomposition (ASVTD) method described in Ruffini et al 2017[14], which takes as input a table where each row corresponds to a patient and each column to an observed variable on patients, such as a diagnostic, a clinical result, demographics such sex and age, etc. plus a number of k desired clusters. This results in the description of the k clusters found, where each cluster is described by the average value of each variable in it. Each patient (in the dataset, or newly arriving patients) can then be assigned to the most-aligned cluster.

This method produces clusters based on logical weight of given attributes. Compared to distance- or similarity-based clustering methods (k-means, k-medoids or PAM, dendograms), MATE is known to work better in the presence of irrelevant or noisy attributes, and does not require the definition of an a-priori "similarity" function to be used (such as Euclidean distance) [14]. The task of choosing a final number of clusters is left to the user, combining the intuitive meaning of each cluster plus the usual requirement to have a small number of clusters. Additional information about cluster analysis is available in supplementary material.

SPSS for Windows® 25.0 (IBM, USA) was used for non-cluster statistical analysis. For descriptive analysis, frequency and percentage were used for the categorical variables, and mean and standard deviation (SD) or median and interquartile range for continuous variables, when appropriate. For comparative analysis of categorical variables, chi-squared test or Fisher's exact test were used when required. For continuous variables ANOVA or the corresponding non-parametrical test were used when appropriate. Time to event data (time to lung transplant and/or death) were analyzed using Kaplan-Meier survival analysis. A p-value of <0.05 was considered statistically significant. STOBE initiative recommendations were followed [19].

RESULTS

Patient features

Baseline characteristics of the 130 patients enrolled are shown in table 1. The mean age was 69 years old (7.8 SD), and 81% were male. Regarding toxic habits, 72% had smoking exposure of whom 46% had a cumulative

dose associated with a IPF risk factor (≥ 20 pack-years)[20], 12% had a history of alcohol abuse (≥ 3 standard drink/day). 33% of cases were obese and 1.5% underweight. 65% of patients were referred because of respiratory symptoms. Exertional dyspnea was present in 83% of patients at diagnosis, and 65% referred dry cough. Velcro crackles on chest auscultation and clubbing finger were present in 91% and 50% of patients respectively. 46% of patients showed a consistent usual interstitial pneumonia (UIP) pattern in the chest HRCT, 42% probable UIP pattern and 12% indeterminate pattern for UIP. Lung biopsy was performed in 52 cases; 48 surgical biopsy and 4 cryobiopsies (Table 1). Telomere length analysis had been performed on 79 patients with family aggregation or some telomeric clinical sign. Familial aggregation was identified in 28% of cases and telomere shortening was recognized in 18% of patients.

The main comorbidities at diagnosis are shown in table 2. Charlson's comorbidity index was 4.7 (SD 1.7). Cardiovascular risk factors were prevalent, and 39% of cases had at least two factors: arterial hypertension (52%), dyslipidemia (45%) or diabetes mellitus (22%). Symptomatic gastroesophagical reflux disease (GERD) was referred by 45% patients, while the hiatal hernia measured by HRCT was 5% severe and 25% moderate. Emphysema was detected in 33% of patients, but only 11% satisfied the CPFE diagnostic criteria[21]. Heart disease was found in 23% of the participants, most of them in the form of ischemic cardiopathy (15%). Pulmonary arterial hypertension (PAH) was suspected by echocardiography in 32% of patients, but only 6% had PAH by right catheterization and received specific treatment. Sleep study was performed on 29 patients who presented clinical symptoms of obstructive sleep apnea (OSA), of whom 14 were OSA under continuous positive airway pressure (CPAP) treatment and 13 diagnosed with sleep-related hypoxemia (SpO_2 ≤88% for ≥ 5 min) according to International Classification of Sleep Disorders criteria[22].

Patient follow-up and outcomes are depicted in table 3. At the initiation of antifibrotic treatment most patients presented preserved or mild decrease of FVC but severe DLCO deterioration. After 3-year follow-up, 18% of subjects had at least 2 respiratory infections per year in a minimum of 2 years without requiring hospital admission; 22% suffered an acute exacerbation requiring hospital admission. 34% stopped or switched antifibrotic drug due to adverse effects and 5% altered protocol due to IPF progression. 42% of patients didn't show disease progression after 3 years, 32% showed a decline of predicted FVC ≥ 10% and 15% a decline in predicted DLCO ≥ 15% in one year. Lung transplant or death related to IPF progression was observed in 28% of cases (9% and 19% respectively).

IPF clustering

The cluster analysis identified 3 different types of patient groups, aggregating 60, 22 and 48 cases in each group. This clustering grouped the cases by similar disease behavior and patient features, including death, lung transplant and PFS after 3-year follow-up. The characteristics of each cluster are shown in figure 2. Furthermore, values and significance of each variable is exposed in table 4.

Cluster 1 was significantly associated ($p=0.02$) with higher mortality, as shown in Kaplan-Meier progression-free survival curve (Figure 3). 40% of patients in this cluster died or underwent lung transplantation after 3-year follow-up. Median survival time was 113 weeks (interquartile range [IQR] 109). It is remarkable, that 48% of the cluster presented a delay of more than 2 years from the first symptom to the beginning of the antifibrotic treatment. Interestingly, the whole cluster presented tobacco exposure of ≥ 20 pack-years and usual interstitial pneumonia (UIP) pattern in chest HRCT at diagnosis. Additionally, it included nearly all CPFEs (22% of cluster) and more severe DLCO decrease at diagnosis (15%). The highest percentage of moderate-severe hiatal hernia measured by HRCT (43%) was included in this cluster. Finally, this was the only group with low weight (2 patients).

Cluster 2 had the longest progression-free survival and it was predominantly characterized by having less than 2 years delay from the symptoms to the beginning of the antifibrotic treatment, no smoking history and no

clear factor for comorbidity. This cluster has the highest percentage of ILD suspicion due to incidental findings in a radiological study by a non-respiratory cause (50%) or screening in the context of subclinical family aggregation (45%). These patients did not present consistent UIP pattern on chest HRCT and only a minority of cases had ≥ 2 respiratory infections per year.

Cluster 3 showed the highest percentage of disease progression, but with lower mortality than cluster 1. Cluster 3 was characterized by a high rate of metabolic syndrome, including dyslipidemia (56%), obesity (46%), and arterial hypertension (54%). Severe OSA with CPAP treatment was present in 17% of cases. Cardiomyopathy was observed in 29% of cases. Although mean FVC and DLCO were not severely decreased and no consistent UIP pattern was present at diagnosis, moderate or severe dyspnea was referred in 60% of cases and 21% showed a relevant limitation for exercise capacity (less than 350mt in 6MWD). GERD was present in most cases (60%).

DISCUSSION

The cluster analyze of this IPF cohort identifies 3 different types of patients with similar clinical features and disease behavior. Cluster 1 presents the worse 3-year survival rate and involves patients with diagnostic and treatment delay, consistent UIP pattern, smoking history, and emphysema. Although cluster 2 and 3 present a similar prognosis, patient features are different. Cluster 3 includes predominantly patients with obesity and other associated comorbidities such as cardiovascular diseases and OSA syndrome. Different clinical features and comorbidities may impact on quality of life and prognosis of IPF patients [23]. Inclusion of comorbidities in a new multivariable model for predicting the risk of progression (ATORVAN)[2] has improved over previous models such as GAP [8]. Furthermore, distinct phenotypes with different clinical outcomes and healthcare requirements have been suggested using clustering analysis of chronic ILDs [13]. Therefore, clustering IPF by patient features and disease behavior may also help predict outcome and identify patients' needs. This model elucidates associations between clinical data, comorbidities and evolution by clinical clusters.

Diagnostic and treatment delay are the most outstanding factors of cluster 1; 48% of patients had an average wait time of more than 2 years from the first respiratory symptoms to antifibrotic treatment initiation. This diagnostic delay has been described by *Lamas et al,* as an independent variable of mortality, and notes the highest percentage of former smokers in the group with 2-4 year of delay [24]. Interestingly, all patients in cluster 1 were former-smokers. Thus, tobacco is a risk factor in pulmonary fibrosis and emphysema [1], but also it could also be a confounding factor that causes a delay in the assessment of respiratory symptoms. On the other hand, a recent study found the use of inhaled therapy as the most important risk factor for delayed IPF diagnosis [25]. Although cluster 1 had 30% of patients over 75 years old and it is possible that the age may impact on the time to refer patient-symptoms, no significant differences in age between clusters were observed. A high rate of consistent UIP pattern in the chest HRCT at diagnosis has been associated with the delay in IPF diagnosis (28). Similar to Hoyer et al, our study shows a predominance of consistent UIP pattern and lower FVC and DLCO at diagnosis in these patients. Another factor that has reported a poor survival rate is the presence of CPFE [26, 27], which is present in 22% of cases clustered in this group. Furthermore, hiatal hernia is more frequently observed in this group. An increased incidence of hiatal hernia measured by HRCT in IPF [28] and its association to a worse prognosis has been previously described [16]. Regarding the results, the diagnostic delay may also associate low patient weight at diagnosis, which has been identified as a poor prognostic factor[29].

Cluster 2 included a low number of patients with greater survival time (160 weeks on average), longer anti-fibrotic drug treatment time (median of 153 weeks), and a low rate of disease progression (41%), which may

be related to early diagnosis as the main characteristic of this group. The mean time from the onset of respiratory symptoms to the antifibrotic treatment initiation was 48 weeks and none exceeded 2 years. This could be due to the rate of subclinical patients evaluated in the context of incidental findings or familial screening that have been clustered in this group. The familial study could explain the increased significant telomere shortening in this cluster (36%). Although a minority of new diagnosed cases, these patients could be better managed with preventive measures and comprehensive therapeutic approaches [29, 30].

Metabolic syndrome and cardiovascular comorbidities were the main features of cluster 3, which associates a high rate of disease progression. The association between obesity, dyslipidemia, cardiopathy, reduced physical capacity and exertional dyspnea has been well documented [31–33]. Obesity and the high prevalence of severe OSA could explain the higher rates of cardiovascular comorbidities[32, 33]. Cardiovascular risk factors have also been associated with menopause [34]. This cluster included the majority of women in our cohort. The higher prevalence of severe OSA under CPAP treatment could be explained by the predominance of obesity [35]. In this cluster, 72% of sleep study subjects were diagnosed with OSA, which prevalence is disproportionate as is similar to morbid obesity series[36]. It would suggest a possible under-diagnosis of these disorders and the potential need of systematic screening in these types of IPF patients [37–39]. Obesity probably also plays a major role in the higher incidence of GERD, as previously described [40, 41]. At the same time, GERD can be another risk factor for disease progression and acute exacerbations [42].

The number of patients included in the cluster analysis from a single center and the retrospective nature are the main limitations of this pilot study. Another limitation is the inclusion of patients from a broad period of time, which may have had an impact on time to referral, patient management and clinical outcomes. However, only 13 patients were included between 2012 and 2013, when the access to anti-fibrotic treatment was limited and the awareness of disease lower. Cluster analysis should ideally be performed on large multinational cohorts of more than 1,000 patients for identifying as many patient profiles as possible [12, 13]. However, the highlighted disease and patient features at diagnosis associated with disease outcome by using this methodology that integrates all potential risk factors have revealed at least two major points: diagnostic delay and cardiovascular-metabolic comorbidities. These results should be validated and better explored in prospective multicenter studies.

In conclusion, this cluster study helps analyze IPF patients, a population which consistently present a complex variability of features at diagnosis related to the disease, comorbidities and other patient-related conditions, and automatically clusters them depending on similar features and disease behavior. With further work, cluster studies could identify intricate associations invisible without analysis. Therefore, the cluster analysis at diagnosis could identify different groups of IPF patients that would benefit from a better personalized management and therapeutic approach, which would be useful for anticipating patient needs and required resources.

Acknolewdgements

This study has been funded by: Instituto de Salud Carlos III through the grants CM20/00093 (Co-funded by European Social Fund. ESF investing in your future) and PI18/00367 (Co-funded by European Regional Development Fund, ERDF, a way to build Europe); Spanish Society of Pneumology and Thoracic Surgery (SEPAR) grants 631/2018 and 685/2018; Emerging ILD Group of SEPAR grant 005 (Boehringer-Roche); Pneumology Foundation of Catalonia (FUCAP) grant 2019; Spanish Sleep Society (SES) grant 2019. Investigation support BRN-Fundació Ramon Pla Armengol. We thank CERCA Programme / Generalitat de Catalunya for institutional support.

References

1. Travis WD, King TE, Bateman ED, Lynch DA, Capron F, Center D, Colby T V., Cordier JF, DuBois RM,

Galvin J, Grenier P, Hansell DM, Hunninghake GW, Kitaichi M, Müller NL, Myers JL, Nagai S, Nicholson A, Raghu G, Wallaert B, Brambilla CG, Brown KK, Cherniaev AL, Costabel U, Coultas DB, Davis GS, Demedts MG, Douglas WW, Egan J, Eklund AG, et al. American thoracic society/European respiratory society international multidisciplinary consensus classification of the idiopathic interstitial pneumonias. *Am. J. Respir. Crit. Care Med.* [Internet] American Thoracic SocietyNew York, NY; 2002 [cited 2019 Jul 25]. p. 277–304Available from: http://www.atsjournals.org/doi/abs/10.1164/ajrccm.165.2.ats01.

2. Fisher M, Nathan SD, Hill C, Marshall J, Dejonckheere F, Thuresson PO, Maher TM. Predicting Life Expectancy for Pirfenidone in Idiopathic Pulmonary Fibrosis. *J. Manag. care Spec. Pharm.* [Internet] 2017 [cited 2019 Jul 25]; 23: S17–S24Available from: http://www.ncbi.nlm.nih.gov/pubmed/28287347.
3. Sauleda J, Núñez B, Sala E, Soriano JB. Idiopathic Pulmonary Fibrosis: Epidemiology, Natural History, Phenotypes. *Med. Sci. (Basel, Switzerland)* [Internet] Multidisciplinary Digital Publishing Institute (MDPI); 2018 [cited 2019 Jan 13]; 6Available from: http://www.ncbi.nlm.nih.gov/pubmed/30501130.
4. Buendía-Roldán I, Mejía M, Navarro C, Selman M. Idiopathic pulmonary fibrosis: Clinical behavior and aging associated comorbidities. *Respir. Med.* 2017; 129: 46–52.
5. Torrisi SE, Ley B, Kreuter M, Wijsenbeek M, Vittinghoff E, Collard HR, Vancheri C. The added value of comorbidities in predicting survival in idiopathic pulmonary fibrosis: A multicentre observational study. *Eur. Respir. J.* [Internet] 2019; 53Available from: http://dx.doi.org/10.1183/13993003.01587-2018.
6. Caminati A, Lonati C, Cassandro R, Elia D, Pelosi G, Torre O, Zompatori M, Uslenghi E, Harari S. Comorbidities in idiopathic pulmonary fibrosis: An underestimated issue. *Eur. Respir. Rev.* [Internet] 2019; 28: 1–11Available from: http://dx.doi.org/10.1183/16000617.0044-2019.
7. Barratt SL, Morales M, Spiers T, Al Jboor K, Lamb H, Mulholland S, Edwards A, Gunary R, Meek P, Jordan N, Sharp C, Kendall C, Adamali HI. Specialist palliative care, psychology, interstitial lung disease (ILD) multidisciplinary team meeting: A novel model to address palliative care needs. *BMJ Open Respir. Res.* [Internet] 2018 [cited 2019 Jul 23]; 5: e000360Available from: http://www.ncbi.nlm.nih.gov/pubmed/30622718.
8. Ley B, Ryerson CJ, Vittinghoff E, Ryu JH, Tomassetti S, Lee JS, Poletti V, Buccioli M, Elicker BM, Jones KD, King TE, Collard HR. A multidimensional index and staging system for idiopathic pulmonary fibrosis. *Ann. Intern. Med.* [Internet] American College of Physicians; 2012 [cited 2019 Jul 17]; 156: 684–695Available from: http://annals.org/article.aspx?doi=10.7326/0003-4819-156-10-201205150-00004.
9. Raghu G, Remy-Jardin M, Myers JL, Richeldi L, Ryerson CJ, Lederer DJ, Behr J, Cottin V, Danoff SK, Morell F, Flaherty KR, Wells A, Martinez FJ, Azuma A, Bice TJ, Bouros D, Brown KK, Collard HR, Duggal A, Galvin L, Inoue Y, Jenkins RG, Johkoh T, Kazerooni EA, Kitaichi M, Knight SL, Mansour G, Nicholson AG, Pipavath SNJ, Buendía-Roldán I, et al. Diagnosis of Idiopathic Pulmonary Fibrosis. An Official ATS/ERS/JRS/ALAT Clinical Practice Guideline. *Am. J. Respir. Crit. Care Med.* [Internet] 2018 [cited 2018 Nov 4]; 198: e44–e68Available from: https://www.atsjournals.org/doi/10.1164/rccm.201807-1255ST.
10. Kreuter M, Ehlers-Tenenbaum S, Palmowski K, Bruhwyler J, Oltmanns U, Muley T, Heussel CP, Warth A, Kolb M, Herth FJF. Impact of comorbidities on mortality in patients with idiopathic pulmonary fibrosis. Wu M, editor. *PLoS One* [Internet] Public Library of Science; 2016 [cited 2019 Jul 25]; 11: e0151425Available from: http://dx.plos.org/10.1371/journal.pone.0151425.
11. Pedraza-Serrano F, Jiménez-García R, López-de-Andrés A, Hernández-Barrera V, Esteban-Hernández J, Sánchez-Muñoz G, Puente-Maestu L, de-Miguel-Díez J. Comorbidities and risk of mortality among hospitalized patients with idiopathic pulmonary fibrosis in Spain from 2002 to 2014. *Respir. Med.* [Internet] Elsevier Ltd; 2018; 138: 137–143Available from: https://doi.org/10.1016/j.rmed.2018.04.005.
12. Badagliacca R, Rischard F, Papa S, Kubba S, Vanderpool R, Yuan JXJ, Garcia JGN, Airhart S, Poscia R, Pezzuto B, Manzi G, Miotti C, Luongo F, Scoccia G, Sciomer S, Torre R, Fedele F, Vizza CD. Clinical implications of idiopathic pulmonary arterial hypertension phenotypes defined by cluster analysis. *J. Hear. Lung Transplant.* [Internet] Elsevier Inc.; 2020; 39: 310–320Available from: https://doi.org/10.1016/j.healun.2019.12.012.
13. Adegunsoye A, Oldham JM, Chung JH, Montner SM, Lee C, Witt LJ, Stahlbaum D, Bermea RS, Chen LW, Hsu S, Husain AN, Noth I, Vij R, Strek ME, Churpek M. Phenotypic Clusters Predict Outcomes in a Longitudinal Interstitial Lung Disease Cohort. *Chest* [Internet] Elsevier Inc; 2018; 153: 349–360Available from: https://doi.org/10.1016/j.chest.2017.09.026.
14. Ruffini M., Gavaldà R. LE. Clustering Patients with Tensor Decomposition. *Mach. Learn. Healthc. Conf. (MLHC), Boston, August 2017.* [Internet] 2017; Available from: http://proceedings.mlr.press/v68/ruffini17a.html.
15. Raghu G, Rochwerg B, Zhang Y, Garcia CAC, Azuma A, Behr J, Brozek JL, Collard HR, Cunningham W,

Homma S, Johkoh T, Martinez FJ, Myers J, Protzko SL, Richeldi L, Rind D, Selman M, Theodore A, Wells AU, Hoogsteden H, Schünemann HJ, ATS, ERS, JRS. An official ATS/ERS/JRS/ALAT clinical practice guideline: Treatment of idiopathic pulmonary fibrosis: An update of the 2011 clinical practice guideline. *Am. J. Respir. Crit. Care Med.* 2015; 192: e3–e19.

16. Tossier C, Dupin C, Plantier L, Leger J, Flament T, Favelle O, Lecomte T, Diot P, Marchand-Adam S. Hiatal hernia on thoracic computed tomography in pulmonary fibrosis. *Eur. Respir. J.* [Internet] 2016; 48: 833–842Available from: http://dx.doi.org/10.1183/13993003.01796-2015.
17. Planas-Cerezales L, Arias-Salgado EG, Buendia-Roldán I, Montes-Worboys A, López CE, Vicens-Zygmunt V, Hernaiz PL, Sanuy RL, Leiro-Fernandez V, Vilarnau EB, Llinás ES, Sargatal JD, Abellón RP, Selman M, Molina-Molina M. Predictive factors and prognostic effect of telomere shortening in pulmonary fibrosis. *Respirology* [Internet] 2018 [cited 2018 Nov 11]; Available from: http://doi.wiley.com/10.1111/resp.13423.
18. Collard HR, Ryerson CJ, Corte TJ, Jenkins G, Kondoh Y, Lederer DJ, Lee JS, Maher TM, Wells AU, Antoniou KM, Behr J, Brown KK, Cottin V, Flaherty KR, Fukuoka J, Hansell DM, Johkoh T, Kaminski N, Kim DS, Kolb M, Lynch DA, Myers JL, Raghu G, Richeldi L, Taniguchi H, Martinez FJ. Acute exacerbation of idiopathic pulmonary fibrosis an international working group report. *Am. J. Respir. Crit. Care Med.* 2016; 194: 265–275.
19. Chai K-X, Chen Y-Q, Fan P-L, Yang J, Yuan X. Strobe. *Medicine (Baltimore).* 2018; 97: e11775.
20. Raghu G, Collard HR, Egan JJ, Martinez FJ, Behr J, Brown KK, Colby T V, Cordier JF, Flaherty KR, Lasky JA, Lynch DA, Ryu JH, Swigris JJ, Wells AU, Ancochea J, Bouros D, Carvalho C, Costabel U, Ebina M, Hansell DM, Johkoh T, Kim DS, King Jr. TE, Kondoh Y, Myers J, Muller NL, Nicholson AG, Richeldi L, Selman M, Dudden RF, et al. An official ATS/ERS/JRS/ALAT statement: idiopathic pulmonary fibrosis: evidence-based guidelines for diagnosis and management. *Am J Respir Crit Care Med* 2011/04/08. 2011; 183: 788–824.
21. Cottin V, Nunes H, Brillet PY, Delaval P, Devouassaoux G, Tillie-Leblond I, Israel-Biet D, Court-Fortune I, Valeyre D, Cordier JF, Carré P, Chabot F, Chatté G, Coëtmeur D, Crestani B, Dalphin JC, Dietemann A, Gentil B, Humbert M, Lacronique J, Mairesse M, Marchand E, Reynaud-Gaubert M. Combined pulmonary fibrosis and emphysema: A distinct underrecognised entity. *Eur. Respir. J.* 2005; 26: 586–593.
22. Sateia MJ. International Classifi cation of Sleep Disorders-Th ird Edition Highlights and Modifi cations. 2014; .
23. Kreuter M, Ehlers-Tenenbaum S, Palmowski K, Bruhwyler J, Oltmanns U, Muley T, Heussel CP, Warth A, Kolb M, Herth FJF. Impact of Comorbidities on Mortality in Patients with Idiopathic Pulmonary Fibrosis. Wu M, editor. *PLoS One* [Internet] Public Library of Science; 2016 [cited 2019 Jan 13]; 11: e0151425Available from: http://dx.plos.org/10.1371/journal.pone.0151425.
24. Lamas DJ, Kawut SM, Bagiella E, Philip N, Arcasoy SM, Lederer DJ. Delayed access and survival in idiopathic pulmonary fibrosis: A cohort study. *Am. J. Respir. Crit. Care Med.* American Thoracic Society; 2011; 184: 842–847.
25. Hoyer N, Prior TS, Bendstrup E, Wilcke T, Shaker SB. Risk factors for diagnostic delay in idiopathic pulmonary fibrosis. *Respir. Res.* Respiratory Research; 2019; 20: 1–9.
26. Cottin V. The impact of emphysema in pulmonary fibrosis. *Eur. Respir. Rev.* 2013; 22: 153–157.
27. Mejía M, Carrillo G, Rojas-Serrano J, Estrada A, Suárez T, Alonso D, Barrientos E, Gaxiola M, Navarro C, Selman M. Idiopathic pulmonary fibrosis and emphysema: Decreased survival associated with severe pulmonary arterial hypertension. *Chest* American College of Chest Physicians; 2009; 136: 10–15.
28. Noth I, Zangan SM, Soares R V., Forsythe A, Demchuk C, Takahashi SM, Patel SB, Strek ME, Krishnan JA, Patti MG, MacMahon H. Prevalence of hiatal hernia by blinded multidetector CT in patients with idiopathic pulmonary fibrosis. *Eur. Respir. J.* 2012; 39: 344–351.
29. Molina-Molina M, Wijsenbeek M. Comprehensive Care In Pulmonary Fibrosis BARCELONA RESPIRATORY NETWORK. *BRN Rev* [Internet] 2019 [cited 2020 Jan 25]; 5Available from: www.eu-ipff.org.
30. Jouneau S, Kerjouan M, Rousseau C, Lederlin M, Llamas-Guttierez F, De Latour B, Guillot S, Vernhet L, Desrues B, Thibault R. What are the best indicators to assess malnutrition in idiopathic pulmonary fibrosis patients? A cross-sectional study in a referral center. *Nutrition* [Internet] Elsevier Inc.; 2019 [cited 2020 Nov 15]; 62: 115–121Available from: https://pubmed.ncbi.nlm.nih.gov/30878815/.
31. Molina-Molina M, Aburto M, Acosta O, Ancochea J, Rodríguez-Portal JA, Sauleda J, Lines C, Xaubet A. Importance of early diagnosis and treatment in idiopathic pulmonary fibrosis [Internet]. Expert Rev. Respir. Med. Taylor and Francis Ltd; 2018 [cited 2020 Nov 15]. p. 537–539Available from:

https://pubmed.ncbi.nlm.nih.gov/29718749/.
32. Csige I, Ujvárosy D, Szabó Z, Lorincz I, Paragh G, Harangi M, Somodi S, Santulli G. The Impact of Obesity on the Cardiovascular System. *J. Diabetes Res.* 2018; 2018.
33. Poirier P, Giles TD, Bray GA, Hong Y, Stern JS, Pi-Sunyer FX, Eckel RH. Obesity and cardiovascular disease: Pathophysiology, evaluation, and effect of weight loss: An update of the 1997 American Heart Association Scientific Statement on obesity and heart disease from the Obesity Committee of the Council on Nutrition, Physical. *Circulation* 2006; 113: 898–918.
34. Collins P, Webb CM, de Villiers TJ, Stevenson JC, Panay N, Baber RJ. Cardiovascular risk assessment in women – an update. *Climacteric* 2016; 19: 329–336.
35. Hudgel DW, Patel SR, Ahasic AM, Bartlett SJ, Bessesen DH, Coaker MA, Michelle Fiander P, Grunstein RR, Gurubhagavatula I, Kapur VK, Lettieri CJ, Naughton MT, Owens RL, Pepin JLD, Tuomilehto H, Wilson KC. The role of weight management in the treatment of adult obstructive sleep apnea: An official American thoracic society clinical practice guideline. *Am. J. Respir. Crit. Care Med.* 2018; 198: e70–e87.
36. Gasa M, Salord N, Fortuna AM, Mayos M, Vilarrasa N, Dorca J, Montserrat JM, Bonsignore MR, Monasterio C. Obstructive sleep apnoea and metabolic impairment in severe obesity. *Eur. Respir. J.* 2011; 38: 1089–1097.
37. Bosi M, Milioli G, Fanfulla F, Tomassetti S, Ryu JH, Parrino L, Riccardi S, Melpignano A, Vaudano AE, Ravaglia C, Tantalocco P, Rossi A, Poletti V. OSA and Prolonged Oxygen Desaturation During Sleep are Strong Predictors of Poor Outcome in IPF. *Lung* 2017/07/05. 2017; 195: 643–651.
38. Lancaster LH, Mason WR, Parnell JA, Rice TW, Loyd JE, Milstone AP, Collard HR, Malow BA. Obstructive sleep apnea is common in idiopathic pulmonary fibrosis. *Chest* 2009/07/02. 2009; 136: 772–778.
39. Pihtili A, Bingol Z, Kiyan E, Cuhadaroglu C, Issever H, Gulbaran Z. Obstructive sleep apnea is common in patients with interstitial lung disease. *Sleep Breath* 2013/04/09. 2013; 17: 1281–1288.
40. El-Serag HB, Graham DY, Satia JA, Rabeneck L. Obesity is an independent risk factor for GERD symptoms and erosive esophagitis. *Am. J. Gastroenterol.* 2005; 100: 1243–1250.
41. Jacobson B, Somers S, Fuchs C, Kelly C, Camargo C. Association Between Body Mass Index and Gastroesophageal Reflux Symptoms in Both Normal Weight and Overweight Women. *N Engl J Med* [Internet] 2006; 354: 2340–2348Available from: http://www.ncbi.nlm.nih.gov/pubmed/16738270%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2782772.
42. Wang Z, Bonella F, Li W, Boerner EB, Guo Q, Kong X, Zhang X, Costabel U, Kreuter M. Gastroesophageal Reflux Disease in Idiopathic Pulmonary Fibrosis: Uncertainties and Controversies. *Respiration* 2018; .

Table 1. Patient features at diagnosis

	Subjects, n	130
Age, mean (SD)		69 (7.8)
Gender, n (%)	Male	105 (80.8)
BMI, n (%)	<18,5	2 (1.5)
	18,5-24,9	16 (12.3)
	25-29,9	69 (53.1)
	≥30	42 (32.3)
Smoking exposure, n (%)	< 20 pack-years	33 (25.4)
	≥ 20 pack-years	60 (46.2)
Alcohol, n (%)	Active	15 (11.5)
	Former	6 (4.6)
Reason for consultation, n (%)	Respiratory symptoms	85 (65.4)
	Familiar study	10 (7.7)
	Radiological or primary care protocol (without a respiratory clinic)	35 (26.9)
Cough, n (%)	Present	85 (65.4)
Dyspnoea (mMRC), n (%)	0	22 (16.9)
	1	55 (42.3)

	2	42 (32.3)
	3	11 (8.5)
Crackles, n (%)	Present	118 (90.8)
Clubbing fingers, n (%)	Present	65 (50.0)
Thorax HRCT pattern	Definite UIP pattern	60 (46.2)
	Probable UIP pattern	55 (42.3)
	Pattern indeterminate for UIP	15 (11.5)
Biopsy	Cryobiopsy	4 (3.1)
	Surgical biopsy	48 (36.9)
Anathomopathological patterns	Definite UIP	35 (26.9)
	Probable UIP	10 (7.7)
	Possible UIP	6 (4.6)
	Non-representative	1 (0.8)
Severe physiological limitation at diagnosis	FVC <50%, n (%)	2 (1.5)
	DLCO <30%, n (%)	11 (8.4)
	6MWD <350 meters, n (%)	20 (15.4)
Family aggregation, n (%)		36 (27.7)
Telomere shortening, n (%)		23 (17.7)
Antifibrotic treatment, n (%)	Pirfenidone	63 (48.5)
	Nintedanib	67 (51.5)

SD: Standard deviation; BMI: Body mass index; HRCT: High-resolution computed tomography; UIP: Usual interstitial pneumonia; mMRC: Modified medical research council; FVC: Forced vital capacity; DLCO: Diffusing capacity for carbon monoxide; 6MWT: 6-minute walking test.

Table 2. Patient comorbidities at diagnosis

Cardiovascular Risk Factor, n (%)	>1	**51 (39.2)**
	Arterial hypertension	68 (52.3)
	Dyslipidemia	58 (44.6)
	Diabetes mellitus	28 (21.5)
GERD, n (%)	Present	59 (45.4)
Hiatus hernia (TC measure), n (%)	Mild	65 (50.0)
	Moderate	33 (25.4)
	Severe	7 (5.4)
Emphysema, n (%)		**43 (33.1)**
CPFE, n (%)		14 (10.8)
COPD, n (%)	Present	15 (11.5)
Heart disease, n (%)		30 (23.1)
	Valvular	10 (7.7)
	Ischemic	20 (15.4)
	Arrhythmia	9 (6.9)
Echocardiography pulmonary hypertension, n (%)		41 (31.5)
	PAH treated, n (%)	8 (6.2)
Sleep disorder, n (%)		29 (22.3)
	Severe OSA (IAH ≥ 30)	14 (11%)
	Mild or moderate OSA (IAH<30)	8 (6.2%)
	Sleep-related hypoxemia	13 (10%)
Malignant disease, n (%)		16 (12.3)
	Lung	3 (2.3)
	Urogenital	6 (4.6)
	Digestive	5 (3.8)
	Other	2 (1.6)
Depression / Anxiety, n (%)		12 (9.2)
Chronic Kidney failure, n (%)	Grade ≥ 2	11 (8.5)
Neurologic, n (%)		9 (6.9)
	Degenerative	2 (1.5)
	Ischemic	8 (6.2)
Periphery Vasculopathy, n (%)		9 (6.9)
Liver disease, n (%)		8 (6.2)
Peptic ulcer, n (%)		6 (4.6)
Pulmonary embolism, n (%)		3 (2.3)
Charlson's comorbidity index, mean (SD)	index	4.7 (1.7)

GERD: Gastroesophageal reflux disease; CPFE: combined pulmonary fibrosis and emphysema;
COPD: Chronic obstructive pulmonary disease; PAH: Pulmonary arterial hypertension; OSAS: Obstructive sleep apnea syndrome; CPAP: Continuous positive airway pressure

Table 3.

Pulmonary function tests follow-up					
		0 year (N=130)	1 year (N=113)	2 years (N=80)	3 years (N=50)
FVC (mL), mean (SD)		2742 (821)	2738 (807)	2745 (854)	2658 (860)
FVC (%), mean (SD)		82.6 (17.5)	85,2 (19,3)	85.8 (21.2)	81.7 (19.5)
TLC (mL), mean (SD)		4744 (1183)	4526 (1212)	4397 (1179)	4399 (1152)
TLC (%), mean (SD)		78.9 (15.7)	76.3 (15.6)	74.8 (14.7)	73.1 (15.5)
DLCO (%), mean (SD)		50.8 (16.8)	52.9 (17.8)	53.6 (17.4)	52.0 (15.9)
KCO (%), mean (SD)		75.4 (20.7)	78.7 (23.4)	84.0 (27.6)	82.9 (20.8)
6MWD (meters), mean (SD)		429 (88.7)	427 (104.1)	433 (94.1)	437 (94.2)
After 3-year follow-up					
Respiratory infection pattern, n (%) (≥ 2 respiratory infections per year in at least 2 years of the 3 years of follow-up)					23 (17.7)
Acute exacerbation (hospital admission), n (%)					28 (21.5)
Antifibrotic stop or switch due to adverse effects, n (%)					44 (33.8)
Antifibrotic stop or switch due to IPF progression, n (%)					6 (4.6)
Progression-free survival (PFS), n (%)					55 (42.3)
	FVC progression (decrease ≥10%), n (%)				42 (32.3)
	DLCO progression (decrease ≥15%), n (%)				20 (15.4)
	Lung transplant by IPF progression, n (%)				12 (9.2)
	Death by IPF progression, n (%)				25 (19.2)

SD: Standard deviation; FVC: Forced vital capacity; TLC: Total lung capacity; DLCO: Diffusing capacity for carbon monoxide; KCO: Transfer coefficient of the lung for carbon monoxide; 6MWT: 6-minute walking test; IPF: Idiopathic pulmonary fibrosis

Table 4. Clusters differences analysis

	Cluster 1 (N=60) "delayed treatment"	Cluster 2 (N=22) "early diagnosis"	Cluster 3 (N=48) "cardiovascular comorbidity"	p-value
Age, mean (SD)	68.71 (9.32)	70.34 (6.40)	69.47 (6.28)	0.691
Age > 75 years, n (%)	18 (30%)	5 (22.7%)	8 (16.7%)	0.268
Male, n (%)	54 (90%)	20 (90.9%)	31 (64.6%)	0.002*
Death or Transplant at 3 years, n (%)	24 (40%)	4 (18.2%)	9 (18.8%)	0.026*
Survival time in weeks, median (IQR)	113.0 (108.8)	160.5 (132.8)	134.0 (103.3)	0.084
Time from symptoms to diagnosis in weeks, median (IQR)	104.00 (118)	48.00 (18.0)	54.50 (86.0)	0.007*
Respiratory symptoms > 2 years, n (%)	29 (48.3%)	0	14 (29.17%)	<0.001*
Progression at 3 years, n (%)	39 (65%)	9 (40.9%)	27 (56.3%)	0.142
2 or more respiratory infections, n (%)	12 (20%)	1 (4.6%)	10 (20.8%)	0.219
1 or more severe exacerbation, n (%)	15 (25%)	5 (22.7%)	8 (16.7%)	0.557
Charlson index, median (IQR)	4.50 (3.0)	4.00 (1. 0)	5.00 (2.0)	0.668
FVC <50% of predicted, n (%)	1 (1,7%)	1 (4.5%)	0	0.445
DLCO <30% of predicted, n (%)	9 (15.3%)	1 (4.5%)	1 (2.1%)	0.034*
Distance walked in 6MWT <350 mt, mean (SD)	8 (13.3%)	2 (10%)	10 (20.8%)	0.426
Definite HRCT UIP pattern, n (%)	60 (100%)	0	0	<0.001*
Probable HRCT UIP pattern, n (%)	0	14 (63.6%)	41 (85.4%)	<0.001*
CPFE, n (%)	13 (21.7%)	1 (4.54%)	0	<0.001*
Familial pulmonary fibrosis, n (%)	12 (20%)	10 (45.5%)	14 (29.2%)	0.071
Telomere shortening, n (%)	12 (20%)	8 (36.4%)	3 (6.3%)	0.006*
Obesity, n (%)	19 (31.7%)	1 (4.5%)	22 (45.8%)	0.002*
Low Weight, n (%)	2 (3.3%)	0	0	0.656
mMRC dyspnea score of 2-3, n (%)	24 (40%)	0	29 (60.4%)	<0.001*
Smoking exposure (≥ 20 pack-years), n (%)	60 (100%)	0	0	<0.001*
Smoking exposure (< 20 pack-years), n (%)	0	8 (36.4%)	25 (52.1)	<0.001*
OSAS treated with CPAP, n (%)	6 (10%)	0	8 (16.7%)	0.096
Nocturnal hypoxemia, n (%)	7 (11.7%)	1 (4.5%)	5 (10.4%)	0.741
HRCT moderate and severe hiatal hernia, n (%)	26 (43.3%)	1 (4.5%)	13 (27.1%)	0.002*
Cardiopathy, n (%)	15 (25%)	1 (4.5%)	14 (29.2%)	0.067
Pulmonary arterial hypertension, n (%)	18 (30%)	6 (27.3%)	17 (35.4%)	0.746
Antifibrotic treatment stopped or switched, n (%)	20 (33.3%)	10 (45.5%)	17 (35.4%)	0.593
Antifibrotic treatment stopped,n (%)	18 (30)	3 (13.6)	13 (27.1)	0.322
Antifibrotic treatment length in weeks, median (IQR)	90 (99.3)	153 (41.5)	115 (85.8)	0.009*
Radiological findings without	14 (23.3%)	11 (50%)	10 (20.8%)	0.026*

respiratory symptoms, n (%)				

SD: Standard deviation; IQR: Interquartile range; FVC: Forced vital capacity; DLCO: Diffusing capacity for carbon monoxide; 6MWT: 6-minute walking test; UIP: Usual interstitial pneumonia; CPFE: Combined pulmonary fibrosis and emphysema; mMRC: Modified medical research council; OSAS: Obstructive sleep apnea syndrome; CPAP: Continuous positive airway pressure; GERD: Gastroesophageal reflux disease.
Note: the % used is in relation to the cluster concerned

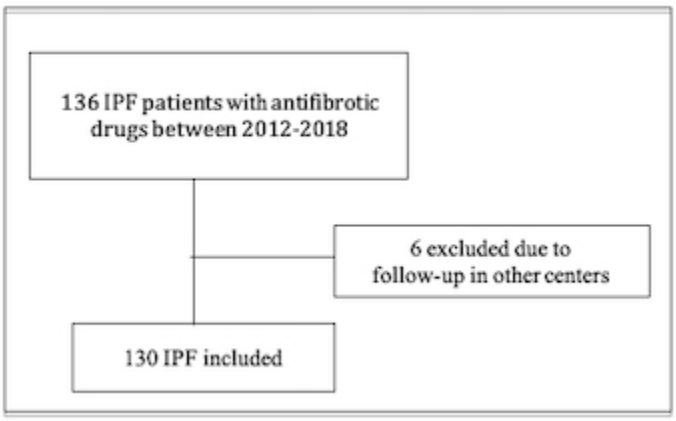
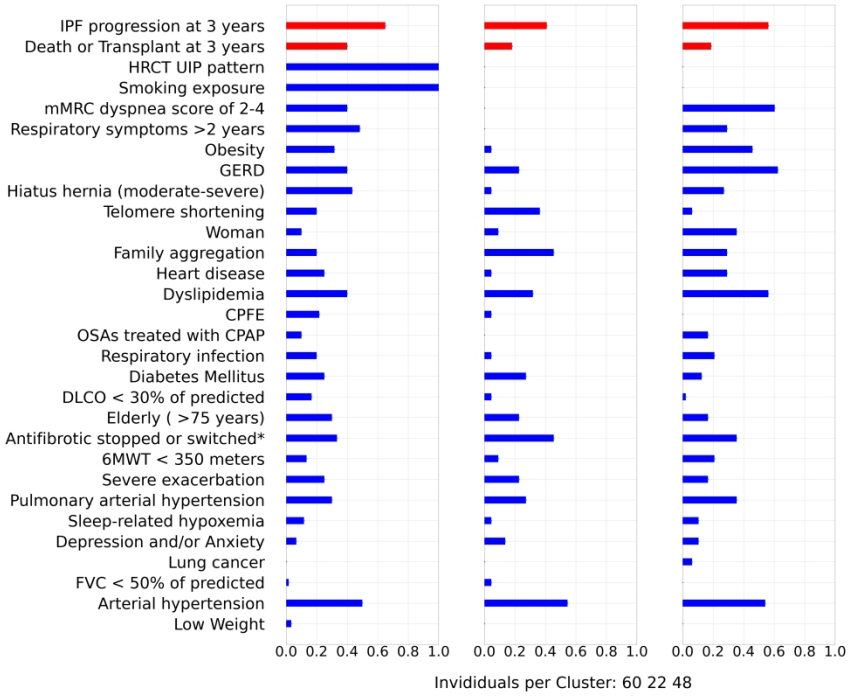


Figure 1. Study flow chart

IPF: Idiopathic pulmonary fibrosis; ILD: Interstitial lung disease.



Caption : Figure 2. Cluster analysis. IPF: Idiopathic pulmonary fibrosis; mMRC: Modified medical research council; FVC: Forced vital capacity; DLCO: Diffusing capacity for carbon monoxide; 6MWT: 6-minute walking test; UIP: Usual interstitial pneumonia; HRCT: High-resolution computed tomography; CPFE: Combined pulmonary fibrosis and emphysema; OSAS: Obstructive sleep apnea syndrome; CPAP: Continuous positive airway pressure; GERD: Gastroesophageal reflux disease. *due to both adverse effect of anti-fibrotic drug treatment and IPF progression.

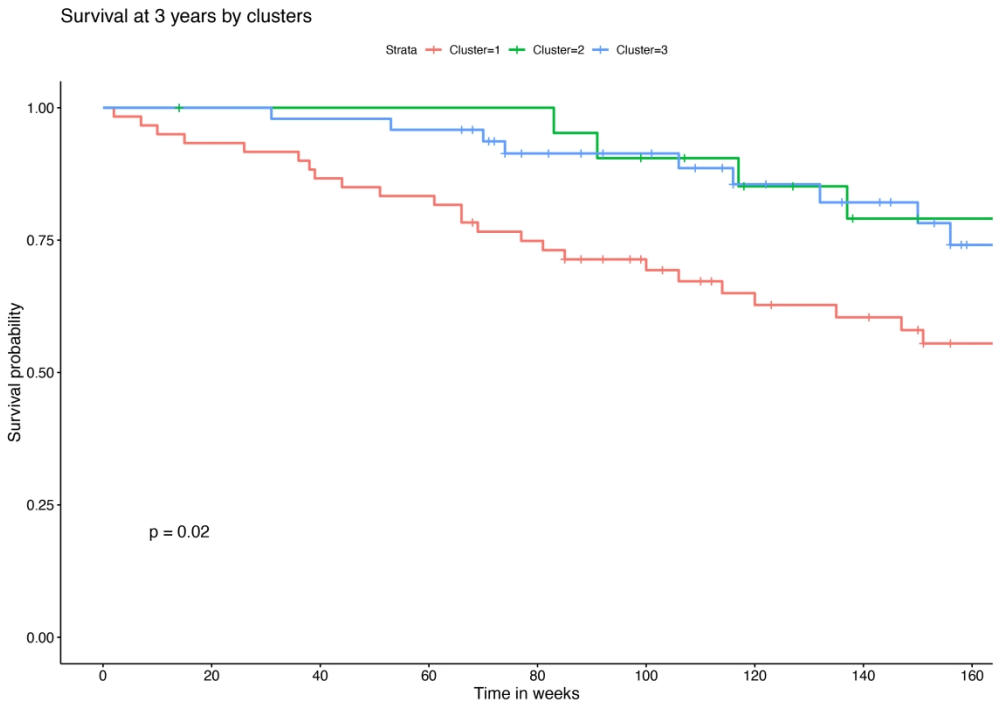


Figure 3. Progression-free survival curve Kaplan-Meier

Supplementary material

We clustered patients using the approximate singular value-based tensor decomposition (ASVTD) method described in Ruffini et al 2017[15]. The method takes as input a table where each row corresponds to a patient and each column to an observed variable on patients, such as a diagnostic, a clinical result, demographics such sex and age, etc. plus a number k of desired clusters. It returns the description of the k clusters found, where each cluster is described by the average value of each observed variable in it.

Most methods used for clustering require a notion of similarity (equivalently, a distance) among patients, and then define clusters so that the intra-cluster similarity is high and the inter-cluster similarity is low; this is the case, for example, for k-means, k-medoids (also known as PAM), and dendogram or hierarchical methods. In contrast, ASVTD takes a probabilistic approach: It assumes that data is generated as a mixture of k unknown populations (the clusters), and finds the descriptions of the k populations whose mixture makes the observed data most likely.

More precisely, ASVTD assumes that there is a set of N observed variables, and a single unobserved or latent variable that takes k possible values. Each unobserved value creates a cluster. Naturally, the observed variables are correlated in arbitrary ways in the data. The main assumption is that, when one fixes the value of the hidden variable (= fixes a cluster), the observed variables become all independent within each cluster, and in particular uncorrelated. The method then finds the partitioning of the instances in k clusters that follows this assumption most closely, that is that makes the observed variables (almost) uncorrelated within each cluster. Then an instance, in the dataset or out of it, can be assigned to the cluster that generates it with the highest probability.

An intuitive explanation of why this strategy makes sense is as follows. Suppose that after partitioning the instances in clusters, two of the observed variables (say, two diagnostics) are still correlated within a cluster. This means that patients in this cluster tend to either have both diagnostics, or to have neither of the two. This in turns means that this cluster can be reasonably subdivided into two clusters: That with patients that have both diagnostics, and that with patients that have neither. Therefore, the clustering is not optimal. Only when all observed variables are independent within every cluster, there is no way of further splitting the clusters more finely.

Compared to similarity-based methods, ASVTD avoids the complicated decision of which distance or similarity function to use, which risks adding a-priori assumptions on the relevant variables. Often, one uses by default the Euclidean distance; this considers equally all attributes, and works badly in high-dimensional data, and especially in the presence of noisy or irrelevant attributes. This does not happen in ASVTD: Irrelevant attributes are not helpful to explain any separation of the data, and are therefore not used in the assignment of instances to clusters. Finally, ASVTD has the potential of creating “all the rest” clusters, collecting the instances that do not fit any of the clear patterns captured by other clusters; this is difficult to do with similarity-based methods.

In ASVTD, the task of choosing a final number of clusters is left to the user. There are mathematical approaches to choosing an optimal number, such as BIC or AIC criteria. In this paper we have used clinical relevance and interpretability as a less formal criterion.

The MATE tool of Amalfi Analytics (www.amalfianalytics.com) has been used in this paper. It implements an extended version of the method in [15] that in particular that can deal with continuous and categorical variables in addition to binary ones.