EUROPEAN RESPIRATORY *journal*

FLAGSHIP SCIENTIFIC JOURNAL OF ERS

Early View

Original article

Machine learning can predict disease manifestations and outcomes in lymphangioleiomyomatosis

Saisakul Chernbumroong, Janice Johnson, Nishant Gupta, Suzanne Miller, Francis X McCormack, Jonathan M Garibaldi, Simon R Johnson

Please cite this article as: Chernbumroong S, Johnson J, Gupta N, *et al*. Machine learning can predict disease manifestations and outcomes in lymphangioleiomyomatosis. *Eur Respir J* 2020; in press (https://doi.org/10.1183/13993003.03036-2020).

Machine learning can predict disease manifestations and outcomes in lymphangioleiomyomatosis.

Saisakul Chernbumroong [1,2], Janice Johnson [3], Nishant Gupta [4], Suzanne Miller [3], Francis X McCormack [4], Jonathan M Garibaldi [2,5], Simon R Johnson [1,3,6]

1. Nottingham Molecular Pathology Node. Nottingham, UK.
2. Advanced Data Analysis Centre. University of Nottingham, UK.
3. Respiratory Medicine and NIHR Biomedical Research Centre. University of Nottingham, UK.
4. Division of Pulmonary, Critical Care and Sleep Medicine, University of Cincinnati, Ohio, USA.
5. School of Computer Science, University of Nottingham, UK.
6. National Centre for Lymphangioleiomyomatosis, Nottingham University Hospitals NHS Trust, UK.

Corresponding Author: Professor Simon Johnson. Division of Respiratory Medicine and Biomedical Research Centre, University of Nottingham, Nottingham Biodiscovery Institute, Science Road, University Park, University of Nottingham. Nottingham NG7 2RD
Email: simon.johnson@nottingham.ac.uk
Telephone: +44 115 8231065

Author contributions: SC and JG performed the machine learning analysis, JJ extracted clinical data, SM performed laboratory analyses, FXM and NG analysed and provided the NHLBI survival data, SRJ conceived the study, obtained the funding, saw the UK patients, performed data analysis and wrote the manuscript. All authors contributed to the final manuscript.

Funding: The study was funded by the Nottingham MRC/EPSRC Molecular Pathology Node and the NIHR Rare disease Translational Research Consortium.

Running head: Machine learning in LAM

Take home message: Using machine learning, simple clinical information from women with LAM can be used to group individuals into clusters. Clusters have differing clinical features, levels of complications and survival and may improve personalised care for LAM.

Word counts: Body text: 2972. Abstract: 240.

Abstract.

Background. LAM is a rare multisystem disease with variable clinical manifestations and differing rates of progression that make management decisions and giving prognostic advice difficult. We used machine learning to identify clusters of associated features which could be used to stratify patients and predict outcomes in individuals.

Patients and methods. Using unsupervised machine learning we generated patient clusters using data from 173 women with LAM from the UK and 186 replication subjects from the NHLBI LAM registry. Prospective outcomes were associated with cluster results.

Results. Two and three-cluster models were developed. A three-cluster model separated a large group of subjects presenting with dyspnoea or pneumothorax from a second cluster with a high prevalence of angiomyolipoma symptoms (p=0.0001) and TSC (p=0.041). The third cluster were older, never presented with dyspnoea or pneumothorax (p=0.0001) and had better lung function. Similar clusters were reproduced in the NHLBI cohort. Assigning patients to clusters predicted prospective outcomes: in a two-cluster model future risk of pneumothorax was 3.3 fold (95% C.I. 1.7-5.6) greater in cluster one than two (p=0.0002). Using the three-cluster model, the need for intervention for angiomyolipoma was lower in clusters two and three than cluster one (p<0.00001). In the NHLBI cohort, the incidence of death or lung transplant was much lower in clusters two and three (p=0.0045).

Conclusions. Machine learning has identified clinically relevant clusters associated with complications and outcome. Assigning individuals to clusters could improve decision making and prognostic information for patients.

Word count: 245

Introduction

Lymphangioleiomyomatosis (LAM) is a rare multisystem disease that occurs both sporadically and in those with TSC[1]. The prevalence of LAM is estimated to be less than 1 per 100 000 women[2] and the diagnosis of an orphan disease is frequently difficult for patients due to feelings of isolation and uncertainty over their prognosis and future disease manifestations[3]. This is particularly true for LAM where both the clinical manifestations and rates of disease progression vary. Although all have lung cysts, only 70% have pneumothorax[4, 5]. Half of women with sporadic LAM and almost all with TSC-LAM have angiomyolipomas, a proportion of which enlarge and are at risk of haemorrhage[6]. Around 20% have significant lymphatic disease[7]. Prognosis can be difficult to predict as some have well preserved lung function long term, whilst others require lung transplantation within a decade of diagnosis.

There are few predictive markers of outcome in LAM. Oestrogen is thought to contribute to disease progression[8-10] and premenopausal status is associated with more rapid loss of lung function[10, 11]. High levels of the lymphangiogenic growth factor, vascular endothelial growth factor type D (VEGF-D) and the presence of bronchodilator reversibility are associated with more rapid loss of FEV_1 in some studies[12, 13] and genetic variants in vitamin D binding protein are associated with shorter survival[14]. Smaller studies have reported other features that are associated with outcome including mode of presentation and initial lung function although all of these associations lack predictive power in individual subjects[15, 16]. Uncertainty around disease progression and complications can worry patients, lead to restrictive lifestyle changes and an unselective approach to management with many given unnecessarily pessimistic advice[17, 18].

We hypothesised that groups of clinical features preferentially cluster together and identifying these associations would improve prediction of complications and outcomes. We used machine learning to associate biological and physiological variables in two national cohorts with the aim of identifying sub-phenotypes within the LAM population that could be used to predict disease manifestations and improve clinical advice.

Methods

The clinical cohorts, variables and analysis are described fully in the on line supplement.

Subjects and clinical data.

The discovery cohort comprised 173 women recruited at the National Centre for LAM in Nottingham UK between 2011 and 2018. All subjects had LAM defined by ATS/JRS criteria[19]. A further 10 were added after the discovery analysis until December 2019. All patients attending the Centre were invited to participate and measurements were made as part of clinical care. At their first visit, which formed the baseline assessment, subjects had CT of the chest, abdomen and pelvis, screening for TSC, lung function, bronchodilator reversibility testing and a six minute walk test according to ERS/ATS standards[20]. CT was used to screen for angiomyolipoma and lymphatic disease, the latter defined as the presence of lymphatic enlargement, chylous pleural effusion or ascites. Review appointments were scheduled according to clinical need and at least annually; complications were recorded, FEV_1 and TL_{CO} were repeated and angiomyolipoma size monitored according to a defined protocol[21]. The East Midlands Research Ethics Committee approved the study (13/EM/0264) and participants gave written informed consent. The replication cohort comprised 186 subjects recruited between 1998 and 2003 to the National Heart Lung and Blood Institute (NHLBI) Registry study on the natural history of LAM[7]. Clinical and serial lung function data were obtained from the National Disease Research Interchange (Philadelphia, USA). All-cause mortality and lung transplantation data for the period until December 2010, prior to the use of rapamycin, were obtained from the United States National Death Index and the United Network for Organ Sharing databases respectively (figure 1).

Cluster assignment was performed using data from the baseline visit (table 1) and outcomes assessed prospectively from this point. Survival is quoted as overall time since diagnosis. Change in lung function was calculated as the slope of all FEV_1 (ΔFEV_1) or TL_{CO} (ΔTL_{CO}) values [22].

Machine learning methodology.

The workflow is summarised in figure 2 and described in detail in the supplementary methods. Briefly, the data set was pre-processed, cleaned and checked for validity. Imputation of missing data was performed using Multiple Imputation Chain Equations (MICE), Random Forest (RF) and MICE with RF. Cluster analysis using multiple algorithms was repeated five times to ensure cluster stability and 42 internal cluster validation schemes applied to determine the optimal number of clusters. We identified the smallest number of variables necessary to classify women with LAM into clusters based on Feature Selection schemes including Recursive Feature Elimination, Correlation-based Feature Detection, Maximum Relevance Minimum Redundant and bivariate statistical tests. Five classification algorithms including Random Forest, Decision Tree, CART, C4.5, C5.0, and Naive Bayes were used to develop models for classifying subjects into clusters. Five-fold cross validation repeated for 10 runs was used when identifying markers and developing classification models. The analysis was carried out using R (https://www.r-project.org/). The clustering algorithms are available at https://github.com/nmpn/lam-stratification.

Statistical analysis.

Data were tested for normality using the Shapiro-Wilk test. Parametric data were analysed using unpaired two-tailed T-test, or one-way ANOVA and non-parametric data using Kruskal-Wallis or Mann-Whitney tests. Categorical data were analysed by Chi Square or Fisher's test. Kolmogorov-Smirnov Tests were used to determine whether two data sets have different distributions. Survival analysis was performed using Kaplan-Meier analysis and Mantel-Cox test. Data were analysed in Microsoft Excel and Graphpad Prism version 7.03.

Results

Cluster model development

Complete demographic, presentation and phenotype data were available for all discovery cohort subjects and treatment, disease activity and oestrogen exposure for greater than 90%. Serum VEGF-D and bronchodilator response data were available for 74 and 61% of subjects respectively (Table 1). Data distribution of missing variables imputed using MICE, RF and MICE+RF did not differ from the original distributions and data imputed from MICE was used (supplementary figure S1).

Two clusters provided optimal separation of factors between groups by majority voting (figure 2 and supplementary table S1). Three clusters also proved clinically useful. Of the five machine learning techniques using fivefold cross validation repeated 10 times, Naïve Bayes delivered the strongest accuracy (0.98, 95% confidence interval 0.9502 - 0.9964), sensitivity (1.0) and specificity (0.96) for cluster assignment and was used henceforth (supplementary table 2 and figure S2). Three classification models were developed, two comprising two clusters and one of three clusters. The initial two-cluster model was based on multiple clustering algorithms, with variables based on feature selection techniques. The alternative two-cluster model used multiple clustering algorithms, with variables based on statistical tests. Whilst both models produced similar groupings, the latter separated subjects using fewer terms, was more effective at predicting complications and is reported henceforth. The three-cluster model was based on hierarchy and Kmeans, with selected variables based on statistics comparing clusters. Subjects were assigned to the cluster for which the output probability was between 0.5 and 1.

Two-cluster model

Thirteen input variables divided subjects into clusters comprising 51 and 49% of the discovery cohort (table 2). The most informative factors discriminating clusters were age at first LAM symptom ($p=7.6x10^{-7}$), age at assessment ($p=4x10^{-14}$), presentation with dyspnoea (p=0.00001), pneumothorax (p=0.00001), angiomyolipoma (p=0.00001) or as a chance finding (p=0.00001), ever experiencing pneumothorax (p=0.00001) or angiomyolipoma (p=0.00017) and baseline TL_{CO} (p=0.0097) (Supplementary figure S2). Cluster one was comprised of younger women with earlier onset disease, predominantly presenting with pneumothorax or angiomyolipoma that had often required

intervention, whereas lymphatic manifestations were uncommon. Subjects in cluster two were on average, 10 years older, tended to present with dyspnoea, had more lymphatic complications and larger defects in gas transfer (lower TL_{CO} and post exercise SaO_2). Pneumothorax was infrequent and although many had angiomyolipoma these seldom required intervention (table 2, supplementary tables S3, S4 and supplementary figure S4).

Three-cluster model

In the three-cluster system, cluster one comprised 69% of subjects who were most likely to present with dyspnoea or pneumothorax and had moderately impaired lung function. Cluster two comprised 22% who very commonly presented with angiomyolipoma related problems, rather than respiratory symptoms, a higher prevalence of TSC and better lung function than cluster one. Cluster three comprised only 9% of subjects and were older at presentation with more recent symptom onset which comprised respiratory symptoms other than breathlessness or pneumothorax, or without LAM symptoms after investigations for other issues. Pneumothorax was very infrequent and lung function almost normal (table 3, figure 3, supplementary figure S3, supplementary tables S5 and S6).

Cluster validation.

To determine if these clusters could be reproduced in other populations, we used subjects recruited in a different country and time period from the discovery cohort. The NHLBI cohort were slightly younger with better lung function than the UK cohort, angiomyolipoma was less common, although other clinical characteristics were similar and age at diagnosis was used in place of age at first symptom. Applying the algorithm without imputation of missing data reproduced both models with a similar level of differentiation other than for angiomyolipoma (figure 4, supplementary tables S7 and S8).

The effect of missing data on cluster assignment was examined by running the clustering algorithm with single factors omitted. Running the three-cluster model using 112 UK subjects for whom all factors were available, was compared with sequential removal of each factor. Omission of factors resulted in misclassifications in a median of 0.7% (range 0-7.1) subjects in cluster one, 5.4% (0-38) in

cluster two and 8.3% (0-17) in cluster three. The chance of misclassification was greater where the original clustering probability was closer to 0.5 than 1 and with omission of factors with the greatest contribution to cluster separation; such as age at first symptom (figure 4, supplementary figures S5 and S6).

Association of clusters with clinical outcomes

To determine if the models could be used to predict outcomes, we examined lung function decline and disease related complications prospectively from the point of cluster assignment and survival from diagnosis. As rapamycin reduces lung function decline, rapamycin treated, and untreated subjects were examined separately. Serial lung function data spanning 54 (SD 36) and 38 (17) months were available for 112 UK and 174 US subjects respectively who had not received rapamycin and for 81 UK subjects treated with rapamycin for a mean of 45 (30) months. There were no significant differences between clusters in rate of loss of FEV_1 or TL_{CO} using either model for untreated or rapamycin treated subjects (figure 5a, supplementary tables S9 and S10).

UK subjects are screened for angiomyolipoma at baseline and tumours monitored using a standardised protocol[21]. Risk of angiomyolipoma intervention was examined irrespective of treatment with rapamycin. Using the two-cluster model, risk of intervention was 0.059 patient-years after assignment to cluster one and 0.025 for cluster two ($p<0.00001$). In the three-cluster model, despite a high prevalence of angiomyolipoma in clusters two and three their need for interventions were significantly lower than in cluster one ($p<0.00001$. Supplementary table S11).

Future risk of pneumothorax was greatest in cluster one using both models in both cohorts (supplementary figure S7). The two-cluster model had the best predictive power where combining all subjects showed the risk of pneumothorax was 3.3-fold (95% C.I. 1.7-5.6) greater in cluster one than two ($p=0.0002$, figure 5b).

Survival and transplant data were available for 166 patients in the NHLBI cohort. Over a mean follow-up of 14 years from cluster assignment and up to 33 years from diagnosis; 38 had required lung transplantation and 14 had died. Time to the combined endpoint of death or transplant was similar

in the two-cluster model (table 5 and supplementary figure S8). In the three-cluster model the incidence of death or transplant was 41.7% in cluster one, zero in cluster two and 4.2 in cluster three (p=0.0045. Figure 5c, supplementary table 12).

Discussion

By applying machine learning to carefully characterised clinical cohorts we have identified groups of related factors which are together associated with outcomes in women with LAM. Whilst clinicians, and indeed patients, have recognised some associations between disease related manifestations, our data for the first time, allow us to quantify the risk of complications, improve prognostic advice and work toward stratified care. Separation into three clusters identifies a large cluster tending to present with pneumothorax or dyspnoea. The second cluster are on average, five years younger with a high prevalence of angiomyolipoma symptoms and TSC. Women in cluster three, whilst comprising only 9% of subjects presented 10-15 years later than clusters one and two with non-classical or no symptoms, didn't experience pneumothorax and tended to have almost normal lung function. Cluster one represents the classic description of women with LAM, presenting in their mid-30s with dyspnoea or pneumothorax and airflow obstruction. Cluster two, where angiomyolipoma haemorrhage or TSC are the first clue to the presence of LAM and respiratory disease is less severe. The third cluster are an increasingly recognised group with milder disease who present at an older age with non-classical symptoms including haemoptysis and cough, or without LAM symptoms. We feel our findings are widely applicable and robust as we were able to independently replicate clusters and although accuracy was reduced somewhat by missing data, the factors required for clustering are available in routine practice. Factors less commonly measured and requiring imputation in the initial analysis, including exertional hypoxaemia, bronchodilator reversibility and VEGF-D were not required for clustering.

The importance of our findings lies in the differences in clinical manifestations, complications and outcomes between clusters. Women with LAM present at varying ages with different symptoms, lung

function and menopausal status. Current guidelines do not give guidance on risk of complications or survival and patients with markedly differing disease may receive similar clinical advice[18, 19, 23]. Applying the methodology described here, could allow clinical advice and decision-making to be improved. Those assigned to clusters two and three presenting in their fifties or later could be reassured that their lifespan is unlikely to be shortened by LAM. The risk of pneumothorax is a common concern[17] and applying the two-cluster model can better quantify this risk with individuals in cluster one having a 10% one year and 43% five year risk of pneumothorax compared with 0 and 15% respectively in cluster two. Such data could be used to improve both patient advice and inform discussions on the need for preventative surgery. Despite a higher prevalence of angiomyolipoma in clusters two and three, the risk of an intervention during follow up is lower than cluster one and the need for surveillance may be less in these groups. This reflects the differing natural history of angiomyolipoma across the clusters: with cluster two and to a lesser extent three, more likely to present with angiomyolipoma and need intervention than cluster one; meaning enlarging and symptomatic tumours have already been treated. The absence of presentation with angiomyolipoma symptoms in cluster one, despite an angiomyolipoma prevalence approaching 50% suggests that angiomyolipoma is often overlooked in this group and makes intervention more likely in these newly identified tumours.

The use of unsupervised machine learning informs us both which variables are important in phenotyping subjects and also understanding the disease. Input variables were chosen for their potential relevance to LAM based on disease manifestations and previous literature. These features included mode and age of presentation, existing clinical manifestations, their severity, oestrogen exposure and pattern of lung physiology. The strongest factors separating clusters being age at first symptom and age at time of assessment. We are unable to say whether clusters represent discreet endotypes: clusters may reflect differences in disease activity with lead-time bias separating subjects presenting earlier due to pneumothorax or angiomyolipoma rather than later with dyspnoea. However, as rate of FEV_1 decline, the best-documented marker of disease activity[9, 10, 24], is similar in all clusters, and clusters have separate disease manifestations suggesting differences in organ

involvement, it seems likely the clusters represent discreet endotypes. In either case, assigning women with LAM to these clusters may be clinically useful. The molecular and cellular processes underlying differences between clusters are not clear and further work examining biomarkers and histologic features within the clusters is required. This initial study shows that machine learning can be applied to the relatively small datasets provided by rare lung diseases using only basic clinical data. Improvements in imaging and biomarker development mean that these variables could be factored into future models which may further improve predictive accuracy.

Our findings are based on two of the largest and best categorised cohorts of women with LAM reported; yet despite using unbiased methodology the study has some limitations. The third cluster in both cohorts comprised a relatively small number of subjects that may have some inbuilt survivor bias. Some variables require further assessment; pre-menopausal status has been associated with accelerated loss of lung function. Menopausal status was not a strong differentiator between clusters and rate of loss of FEV_1 and DL_{CO} were similar between clusters despite differing proportions of pre-menopausal women. Age was a strong determinant of cluster assignment, as menopausal status and age are related, menopausal status may still contribute to some of these differences and should continue to be a factor in clinical decisions. Due to differences in data recording between the UK and US we were unable to reproduce all data, particularly for angiomyolipoma. Since the NHLBI cohort closed, rapamycin has become the standard of care for those with progressive disease[23] and has improved outcomes. How rapamycin affects different clusters and how clustering may inform the decision to use rapamycin should be studied prospectively; including using data from the ongoing Multicenter Interventional Lymphangioleiomyomatosis Early Disease Trial (NCT03150914). Our study was not designed to predict need for therapy, however it could be argued that those in cluster one should already be considered for early treatment with mTOR inhibitors to prevent further loss of lung function.

In conclusion, we have used machine learning techniques to stratify women with LAM into clusters using simple clinical data. The method has the potential to improve advice on disease trajectory,

complications and screening. Further prospective studies are warranted to determine if this can be translated to improve management for women with LAM.

Acknowledgements. We are grateful to the original NHLBI cohort investigators, the women with LAM who contributed to the study and Anne Tattersfield for critical reading of the manuscript.

Table 1. Disease related variables captured in discovery and replication cohorts.

Cohort		**UK (n=173)**		**NHLBI (n=186)**	
Variable	Data type	Missing (%)	Mean (SD) or % present	Missing (%)	Mean (SD) or % present
Demographic					
Age (years)	Continuous	0	48.5 (11.8)	0	45.0 (9.3)
Age 1st symptom (years)	Continuous	0	35.7 (11.5)	-	NA
Age at diagnosis (years)	Continuous	-	NA	0	40.7 (9.5)
Disease duration (years)	Continuous	0	12.8 (10.2)	-	NA
Time since diagnosis (years)	Continuous	-	NA	0	4.4 (4.24)
Body mass index (kg/m^2)	Continuous	0	26.2 (6.3)	-	NA
First symptom *					
Dyspnoea (%)	Categorical	0	39	0	48
Pneumothorax (%)	Categorical	0	27	0	33
Other respiratory (%)	Categorical	0	9	0	7
Angiomyolipoma (%)	Categorical	0	15	0	4
Other non-respiratory (%)	Categorical	0	3	0	2
Screened (%)	Categorical	0	3	0	5
None (%)	Categorical	0	4	0	1
Phenotype †					
Tuberous sclerosis present (%)	Categorical	0	21	0	10
Ever had angiomyolipoma (%)	Categorical	0	64	0	18
Lymphatic disease (%)	Categorical	0	17	0	17
Ever had pneumothorax (%)	Categorical	0	44	0	53
Oestrogen exposure					
Number of children	Continuous	2.3	0.96 (1.1)	-	NA
Post menopause (%)	Categorical	1.1	34	0	15
Disease activity markers					
Surgery for pneumothorax (%)	Categorical	0.6	34	0	23
Intervention for angiomyolipoma (%)	Categorical	1.1	37	0	15
Serum VEGF-D (pg/ml)	Continuous	26	1407 (1392)	-	NA
Physiology at enrolment					
FEV_1 (% predicted)	Continuous	4.6	68.3 (26)	0	74.2 (25)
TL_{CO} (% predicted)	Continuous	6.9	52.3 (19.8)	1.6	57.4 (22)
$\%FEV_1/\%TL_{CO}$	Continuous	7.5	1.37 (0.44)	1.6	1.40 (0.41)
Post walk SaO_2 (%)	Continuous	10	87.9 (6.8)	-	NA
Positive bronchodilator response (%)	Categorical	39	62	1.1	38
Treatment at enrolment					
On rapamycin (%)	Categorical	0.5	52	0	0
On oxygen (%)	Categorical	0.5	23	-	NA

'Disease duration' is defined as time from first LAM symptom to baseline study assessment. *, The first recorded symptom of LAM. Only one of the group for each subject. 'Other respiratory' is any respiratory symptom other than dyspnoea or pneumothorax. 'Other non-respiratory' is any non-respiratory symptom other than angiomyolipoma. †, ever experienced by subject, any combination may be present. NA, not available for this cohort.

Table 2. Discriminating features of the two-cluster model.

Factor	Cluster 1	Cluster 2	Mean diff.	p
n (%)	97 (51)	86 (49)		
Demographic *				
Age at assessment (yrs)	46.6 (11)	54.8 (10.6)	-8.2	$7.6x10^{-7}$
Age 1st symptom (yrs)	31.9 (9.8)	44.4 (10.6)	-12.4	$4x10^{-14}$
Disease duration (months)	143 (120)	90 (84.7)	52	0.00083
BMI (kg/m^2)	24.6 (5)	27.4 (6.9)	-2.7	0.002
VEGF-D (pg /ml)	1319 (1320)	1370 (1328)	-51	0.801
Presenting symptom †				
Dyspnoea	4	49	-45	0.00001
Pneumothorax	54	0	54	0.00001
Other respiratory	3	14	-11	0.011
Angiomyolipoma	32	5	27	0.00001
Screened	2	1	1	0.56
Chance finding	0	9	-9	0.009
Phenotype †				
Ever had pneumothorax	68	16	52	0.00001
Ever had angiomyolipoma	69	43	26	0.00017
Lymphatic disease	13	16	-3	0.546
TSC	17	8	9	0.054
Lung function *				
FEV_1 (% predicted)	72.7 (22.0)	68.4 (26.8)	4.3	0.24
TL_{CO} (% predicted)	58.8 (16.8)	51.5 (20.3)	7.3	0.0097
6 minute walk distance (m)	501 (127)	457 (136)	43	0.103
Post walk saturation (%)	89.1 (6.8)	87.7 (6.9)	1.4	0.268
Bronchodilator reversibility (%)	7.5 (7.5)	10.9 (10.8)	-3.4	0.126

* Mean value (standard deviation) compared by unpaired 2 tail t-test. † Percentage of cohort with this feature present compared by chi square test.

Table 3. Discriminating factors of the three-cluster model.

	Cluster 1	**Cluster 2**	**Cluster 3**	**p**
n (%)	127 (69)	39 (22)	17 (9)	
Demographic *				
Age at assessment (yrs)	50. 4 (11.4)	45.9 (10.4)	60.2 (9.9)	<0.0001
Age 1st symptom (yrs)	37.4 (10.8)	32.0 (10.5)	52.8 (10.1)	<0.0001
Disease duration (months)	120 (108)	136 (113)	59 (61)	0.043
BMI (kg/m^2)	26 (6.0)	26 (6.2)	28 (6.8)	0.2
VEGF-D (pg /ml)	1385 (1431)	1286 (1099)	1141 (816)	0.26
Presenting symptom †				
Dyspnoea	41.7	0	0	0.0001
Pneumothorax	42.5	0	0	0.0001
Other respiratory	8.7	2.5	29.4	0.0057
Angiomyolipoma	0	89.7	11.8	0.0001
Screened	0.8	5.1	0	0.156
Chance finding	2.4	2.5	29	0.0001
Phenotype †				
Ever had pneumothorax	58.3	25.6	0	0.0001
Ever had angiomyolipoma	49.6	97.4	64.7	0.0001
Lymphatic disease	17.3	5.12	29.4	0.051
TSC	11.0	25.6	5.9	0.041
Lung function *				
FEV_1 (% predicted)	64.0 (23.4)	79.6 (26.9)	90.7 (19.0)	<0.0001
TL_{CO} (% predicted)	50.5 (19.9)	62.7 (17.3)	67.0 (10.2)	<0.0001
6 minute walk distance (m)	470 (145)	499 (112)	521 (52)	0.44
Post walk saturation (%)	86.7 (7.1)	90.8 (5.6)	93.4 (2.8)	0.0006
Bronchodilator reversibility (%)	11.1 (10.4)	5.8 (6.2)	5.5 (5.5)	0.066

* mean (+/-SD), analysed by one way ANOVA. † percentage of cohort, analysed by chi square test.

Figure legends

Figure 1. Enrolment and data available in cohorts studied. Women with LAM were recruited from the UK LAM Centre (UK) and the National Heart, Lung and Blood Institute LAM registry in the USA (NHLBI). Not all data were available for all subjects for all endpoints. Exact numbers are specified in the individual analyses.

Figure 2. Study workflow, data identification and separation of features into two clusters. (a) Summary workflow of data processing and analysis. The data set was pre-processed which involved data cleaning and data validity checking. Missing data were imputed using Multiple Imputation Chain Equation (MICE), Random Forest (RF), and MICE + RF. Data were transformed from numerical and categorical variables for clustering analysis using Principal Component Analysis (PCA) with Multiple Correspondent Analysis (MCA) and Gower's distance. Optimal number of cluster identification was performed then internal cluster validity indexes. Gap statistics with bootstrapping were used to determine cluster validity. Cluster analysis using four algorithms and classification models developed using by Recursive Feature Elimination followed by the classification algorithms Naïve Bayes, Random Forest (RF) and Nearest Neighbour. Full details are given in the supplementary methods **(b)** Inertia gain plot measuring the degree of homogeneity between the data associated with a cluster using hierarchical + Kmeans methods. Division of the data into two and three clusters gives good separation. **(c)** Cluster dendrogram showing separation between the three clusters using hierarchical clustering + Kmeans. **(d)** Principal component analysis showing separation of subjects into three clusters.

Figure 3. Features of the three-cluster model. (a) Distribution of age, age at first symptom, percent predicted FEV_1 and TL_{CO} at baseline and hypoxia during exertion in the three-cluster model. **(b)** Representative subjects from clusters one, two and three. Showing at baseline age, presenting symptom, CT images of the chest, abdomen and lung function. Cluster one subject presented age 36 with pneumothorax (grey arrow). Cluster two presented with ruptured angiomyolipoma requiring

embolisation (black arrow). Cluster three subject was diagnosed after a lymphatic mass (white arrow) was detected during a CT scan was performed for another indication.

Figure 4. Cluster validation analyses. (a) Comparison of variable distribution in the UK and NHLBI Cohorts for the three-cluster model. Clusters are represented by the percentage of positive subjects for each variable within that cluster in the two cohorts. *presenting symptom. †feature ever present. **(b)** Effect of missing data upon cluster assignment. 112 subjects from the UK cohort with complete data were assigned to clusters and then reassigned with each variable removed in turn. The heatmap is red for correctly assigned subjects (columns) and tan when omission of that variable (rows) led to mis-assignment to cluster one, purple to cluster two and yellow to cluster three. Subjects for each cluster ranked according to strength of assignment (posterior prediction) to the cluster from 1 (strong) to 0.5 (weak) left to right along the y axis.

Figure 5. Prospective clinical outcomes stratified by cluster. (a) Rate of change of FEV_1 and TL_{CO} (ΔFEV_1 and ΔTL_{CO}) for subjects in the UK and NHLBI cohorts combined who were not being treated with rapamycin stratified using the two and three-cluster models. Values within bars are the number of subjects with lung function data available for analysis. None of the differences between clusters in the models was significant. **(b)** Kaplan Meier analysis of the prospective risk of pneumothorax following cluster assignment in the UK and NHLBI cohorts combined for the two-cluster model. Those in cluster one have a 3.3 fold higher risk of pneumothorax, independent of prior treatment for pneumothorax compared with those in cluster two. **(c)** Kaplan Meier analysis of the combined risk of death or need for lung transplantation since diagnosis in the NHLBI cohort stratified using the three-cluster model.

References

1. Johnson SR, Taveira-DaSilva AM, Moss J. Lymphangioleiomyomatosis. *Clinics in Chest Medicine* 2016: 37(3): 389-403.
2. Harknett EC, Chang WYC, Byrnes S, Johnson J, Lazor R, Cohen MM, Gray B, Geiling S, Telford H, Tattersfield AE, Hubbard RB, Johnson SR. Regional and National Variability Suggests Underestimation of Prevalence of Lymphangioleiomyomatosis. *Quarterly Journal of Medicine* 2011: 104(11): 971-979.
3. Carel H, Johnson S, Gamble L. Living with lymphangioleiomyomatosis. *BMJ* 2010: 340(mar12_1): c848-.
4. Taveira-DaSilva AM, Steagall WK, Rabel A, Hathaway O, Harari S, Cassandro R, Stylianou M, Moss J. Reversible airflow obstruction in lymphangioleiomyomatosis. *CHEST Journal* 2009: 136(6): 1596-1603.
5. Johnson J, Johnson SR. Cross-sectional study of reversible airway obstruction in LAM: better evidence is needed for bronchodilator and inhaled steroid use. *Thorax* 2019: thoraxjnl-2019-213338.
6. Yeoh Z, Navaratnam V, Bhatt R, McCafferty I, Hubbard R, Johnson S. Natural history of angiomyolipoma in lymphangioleiomyomatosis: implications for screening and surveillance. *Orphanet Journal of Rare Diseases* 2014: 9(1): 151.
7. Ryu JH, Moss J, Beck GJ, Lee J-C, Brown KK, Chapman JT, Finlay GA, Olson EJ, Ruoss SJ, Maurer JR, Raffin TA, Peavy HH, McCarthy K, Taveira-DaSilva A, McCormack FX, Avila NA, DeCastro RM, Jacobs SS, Stylianou M, Fanburg BL, for the NHLBI LAM Registry Group. The NHLBI Lymphangioleiomyomatosis Registry: Characteristics of 230 Patients at Enrollment. *Am J Respir Crit Care Med* 2006: 173(1): 105-111.
8. Johnson SR, Tattersfield AE. Clinical experience of lymphangioleiomyomatosis in the UK. *Thorax* 2000: 55(12): 1052-1057.

9. Taveira-DaSilva AM, Stylianou MP, Hedin CJ, Hathaway O, Moss J. Decline in Lung Function in Patients With Lymphangioleiomyomatosis Treated With or Without Progesterone. *Chest* 2004: 126(6): 1867-1874.
10. Johnson SR, Tattersfield AE. Decline in lung function in lymphangioleiomyomatosis: relation to menopause and progesterone treatment. *Am J Respir Crit Care Med* 1999: 160(2): 628-633.
11. Gupta N, Lee H-S, Young LR, Strange C, Moss J, Singer LG, Nakata K, Barker AF, Chapman JT, Brantly ML, Stocks JM, Brown KK, Lynch JP, Goldberg HJ, Downey GP, Taveira-DaSilva AM, Krischer JP, Setchell K, Trapnell BC, Inoue Y, McCormack FX. Analysis of the MILES Cohort Reveals Determinants of Disease Progression and Treatment Response in Lymphangioleiomyomatosis. *European Respiratory Journal* 2019: 1802066. 2019 Feb;155(2):288-96
12. Young LR, Lee H-S, Inoue Y, Moss J, Singer LG, Strange C, Nakata K, Barker AF, Chapman JT, Brantly ML, Stocks JM, Brown KK, Lynch JP, Goldberg HJ, Downey GP, Swigris JJ, Taveira-DaSilva AM, Krischer JP, Trapnell BC, McCormack FX. Serum VEGF-D concentration as a biomarker of lymphangioleiomyomatosis severity and treatment response: a prospective analysis of the Multicenter International Lymphangioleiomyomatosis Efficacy of Sirolimus (MILES) trial. *The Lancet Respiratory Medicine* 2013.Aug; 1(6):445-52
13. Le K, Steagall WK, Stylianou M, Pacheco-Rodriguez G, Darling TN, Vaughan M, Moss J. Effect of beta-agonists on LAM progression and treatment. *Proceedings of the National Academy of Sciences* 2018: 115(5): E944.
14. Miller S, Coveney C, Johnson J, Farmaki A-E, Gupta N, Tobin MD, Wain LV, McCormack FX, Boocock DJ, Johnson SR. The Vitamin D Binding Protein axis modifies disease severity in Lymphangioleiomyomatosis. *European Respiratory Journal* 2018.Nov 1;52(5)1800951
15. Lazor R, Valeyre D, Lacronique J, Wallaert B, Urban T, Cordier JF. Low initial KCO predicts rapid FEV1 decline in pulmonary lymphangioleiomyomatosis. *Respiratory medicine* 2004: 98(6): 536-541.
16. Cohen MM, Pollock-BarZiv S, Johnson SR. Emerging clinical picture of lymphangioleiomyomatosis. *Thorax* 2005: 60(10): 875-879.

17. Young LR, Almoosa KF, Pollock-BarZiv S, Coutinho M, McCormack FX, Sahn SA. Patient Perspectives on Management of Pneumothorax in Lymphangioleiomyomatosis 10.1378/chest.129.5.1267. *Chest* 2006: 129(5): 1267-1273.

18. Johnson SR, Cordier JF, Lazor R, Cottin V, Costabel U, Harari S, Reynaud-Gaubert M, Boehler A, Brauner M, Popper H, Bonetti F, Kingswood C, the Review Panel of the ERSLAMTF. European Respiratory Society guidelines for the diagnosis and management of lymphangioleiomyomatosis. *The European respiratory journal* 2010: 35(1): 14-26.

19. Gupta N, Finlay GA, Kotloff RM, Strange C, Wilson KC, Young LR, Taveira-DaSilva AM, Johnson SR, Cottin V, Sahn SA, Ryu JH, Seyama K, Inoue Y, Downey GP, Han MK, Colby TV, Wikenheiser-Brokamp KA, Meyer CA, Smith K, Moss J, McCormack FX. Lymphangioleiomyomatosis Diagnosis and Management: High-Resolution Chest Computed Tomography, Transbronchial Lung Biopsy, and Pleural Disease Management. An Official American Thoracic Society/Japanese Respiratory Society Clinical Practice Guideline. *American Journal of Respiratory and Critical Care Medicine* 2017: 196(10): 1337-1348.

20. Miller MR, Crapo R, Hankinson J, Brusasco V, Burgos F, Casaburi R, Coates A, Enright P, Grinten CPMvd, Gustafsson P, Jensen R, Johnson DC, MacIntyre N, McKay R, Navajas D, Pedersen OF, Pellegrino R, Viegi G, Wanger J. General considerations for lung function testing. *European Respiratory Journal* 2005: 26(1): 153.

21. Bee J, Bhatt R, McCafferty I, Johnson S. Audit, research and guideline update: A 4-year prospective evaluation of protocols to improve clinical outcomes for patients with lymphangioleiomyomatosis in a national clinical centre. *Thorax* 2015; 70:1204

22. Bee J, Fuller S, Miller S, Johnson SR. Lung function response and side effects to rapamycin for lymphangioleiomyomatosis: a prospective national cohort study. *Thorax* 2018: 73(4): 369.

23. McCormack FX, Gupta N, Finlay GR, Young LR, Taveira-DaSilva AM, Glasgow CG, Steagall WK, Johnson SR, Sahn SA, Ryu JH, Strange C, Seyama K, Sullivan EJ, Kotloff RM, Downey GP, Chapman JT, Han MK, D'Armiento JM, Inoue Y, Henske EP, Bissler JJ, Colby TV, Kinder BW, Wikenheiser-Brokamp KA, Brown KK, Cordier JF, Meyer C, Cottin V, Brozek JL, Smith K, Wilson KC, Moss J.

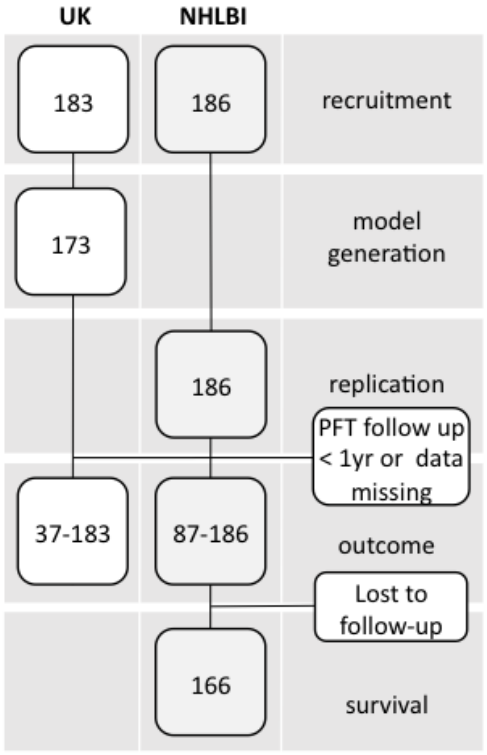
Official American Thoracic Society/Japanese Respiratory Society Clinical Practice Guidelines: Lymphangioleiomyomatosis Diagnosis and Management. *American Journal of Respiratory and Critical Care Medicine* 2016: 194(6): 748-761.

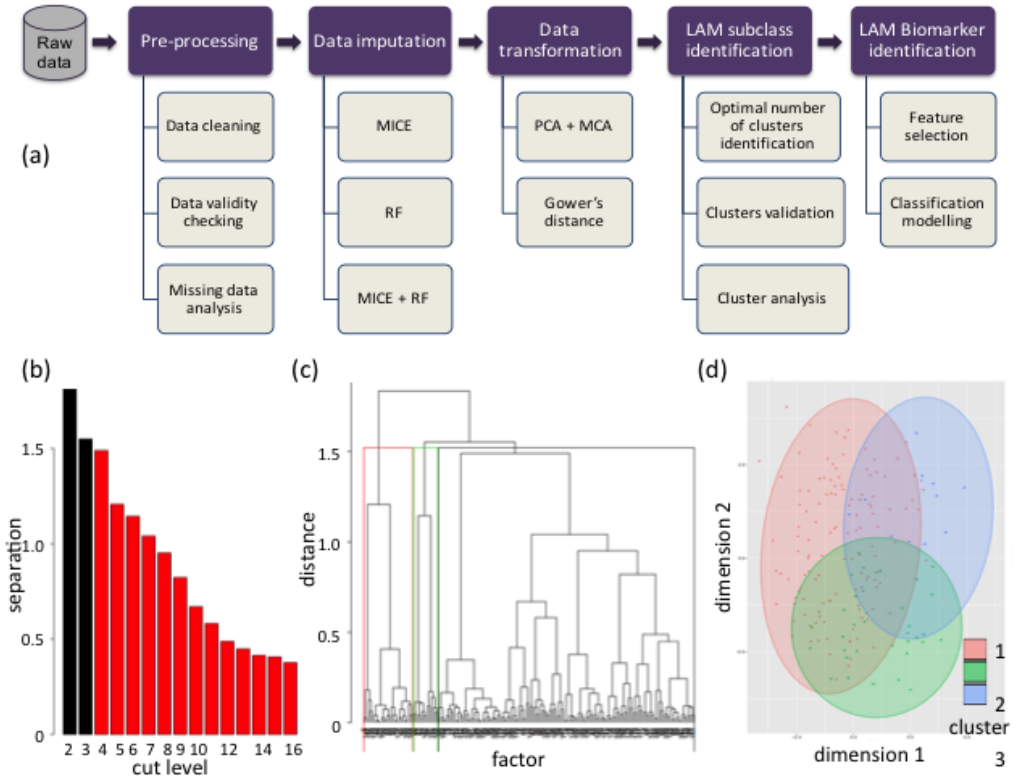
24. McCormack FX, Inoue Y, Moss J, Singer LG, Strange C, Nakata K, Barker AF, Chapman JT, Brantly ML, Stocks JM, Brown KK, Lynch JP, Goldberg HJ, Young LR, Kinder BW, Downey GP, Sullivan EJ, Colby TV, McKay RT, Cohen MM, Korbee L, Taveira-DaSilva AM, Lee H-S, Krischer JP, Trapnell BC. Efficacy and Safety of Sirolimus in Lymphangioleiomyomatosis. *New England Journal of Medicine* 2011: 364: 1595-1606.

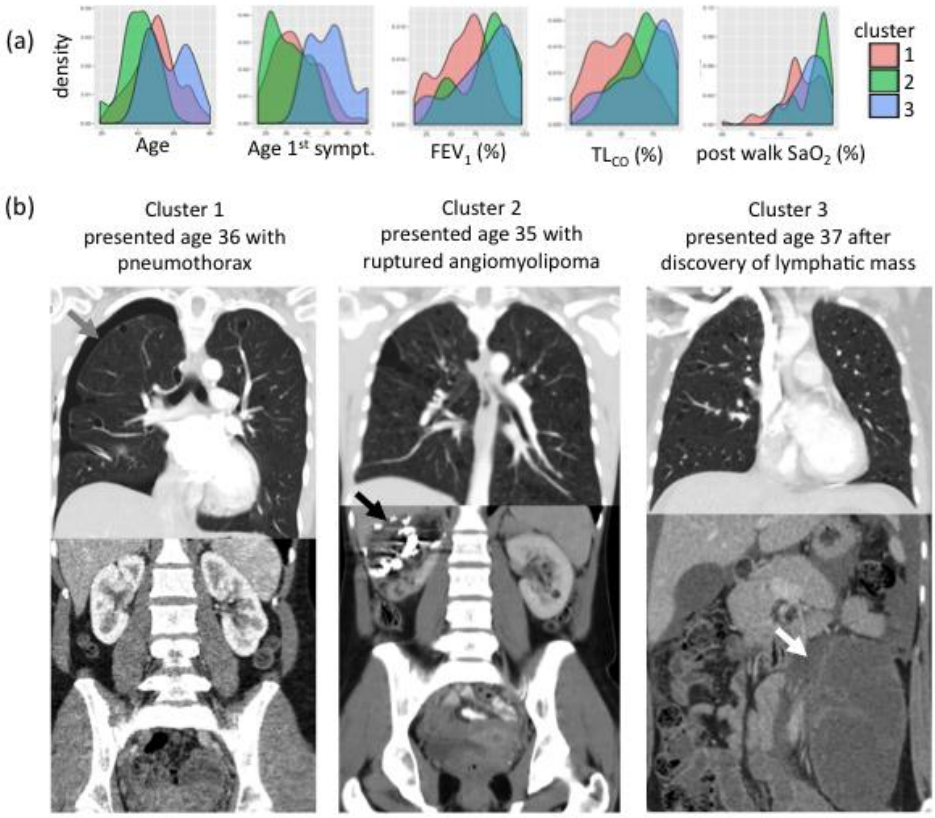
25. Miller S, Stewart ID, Clements D, Soomro I, Babaei-Jadidi R, Johnson SR. Evolution of lung pathology in lymphangioleiomyomatosis: associations with disease course and treatment response. *The journal of pathology Clinical research* 2020.;6:215-26

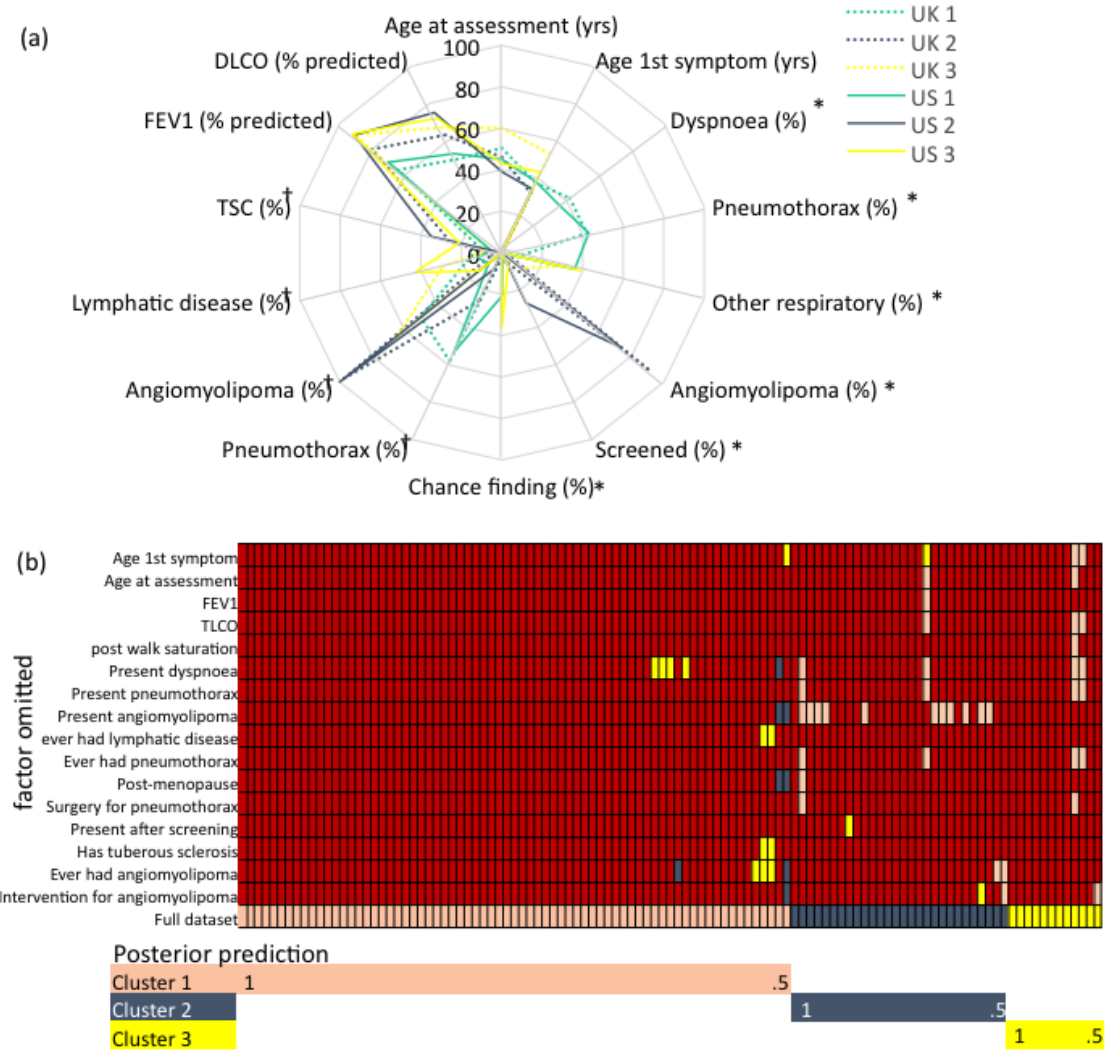
26. Osterburg AR, Nelson RL, Yaniv BZ, Foot R, Donica WRF, Nashu MA, Liu H, Wikenheiser-Brokamp KA, Moss J, Gupta N, McCormack FX, Borchers MT. NK cell activating receptor ligand expression in lymphangioleiomyomatosis is associated with lung function decline. *JCI insight* 2016: 1(16).

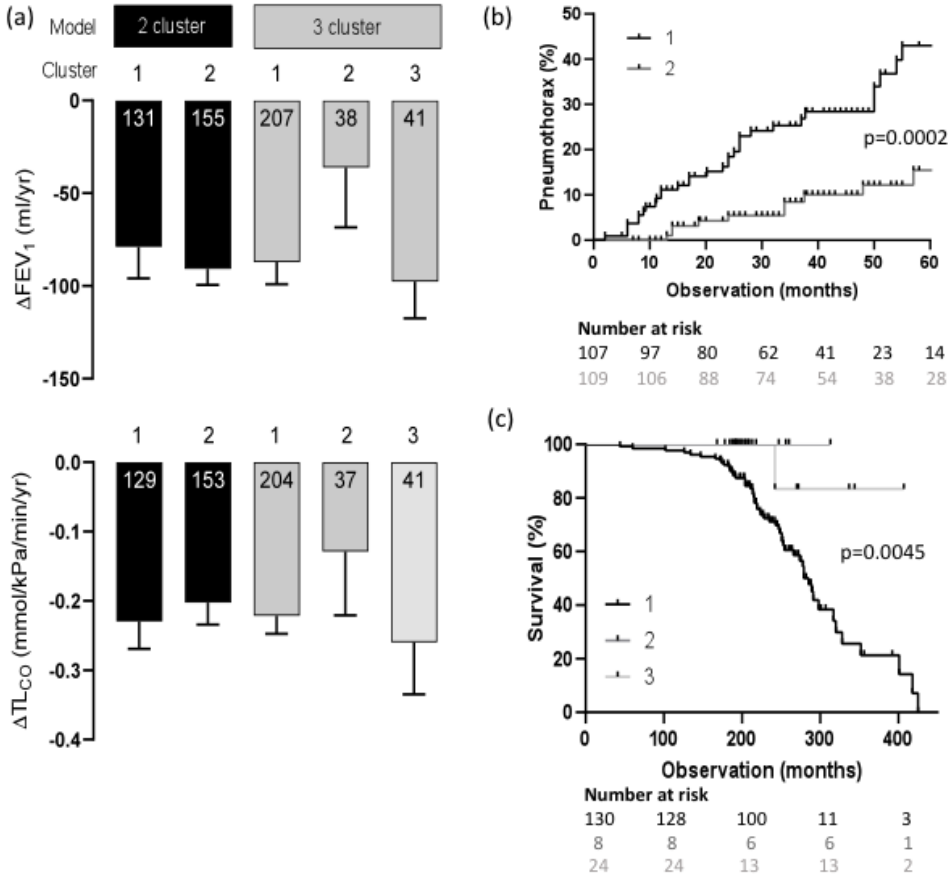
27. Lamattina AM, Poli S, Kidambi P, Bagwe S, Courtwright A, Louis PH, Shrestha S, Stump B, Goldberg HJ, Thiele EA, Rosas I, Henske EP, El-Chemaly S. Serum endostatin levels are associated with diffusion capacity and with tuberous sclerosis- associated lymphangioleiomyomatosis. *Orphanet Journal of Rare Diseases* 2019: 14(1): 72.


UK
NHLBI
183
186
recruitment
173
model generation
186
replication
PFT follow up < 1yr or data missing
37-183
87-186
outcome
Lost to follow-up
166
survival


(a)
Raw data
Pre-processing
Data cleaning
Data validity checking
Missing data analysis
Data imputation
MICE
RF
MICE + RF
Data transformation
PCA + MCA
Gower's distance
LAM subclass identification
Optimal number of clusters identification
Clusters validation
Cluster analysis
LAM Biomarker identification
Feature selection
Classification modelling
(b)
separation
1.5
1.0
0.5
0
2 3 4 5 6 7 8 9 10 12 14 16
cut level
(c)
distance
1.5
1.0
0.5
0
factor
(d)
dimension 2
dimension 1
1
2
3
cluster


(a)
density
Age
Age 1st sympt.
FEV1 (%)
TLCO (%)
post walk SaO2 (%)
cluster
1
2
3
(b)
Cluster 1
presented age 36 with
pneumothorax
Cluster 2
presented age 35 with
ruptured angiomyolipoma
Cluster 3
presented age 37 after
discovery of lymphatic mass


(a)
Age at assessment (yrs)
Age 1st symptom (yrs)
Dyspnoea (%) *
Pneumothorax (%) *
Other respiratory (%) *
Angiomyolipoma (%) *
Screened (%) *
Chance finding (%)*
Pneumothorax (%)†
Angiomyolipoma (%)†
Lymphatic disease (%)†
TSC (%)†
FEV1 (% predicted)
DLCO (% predicted)
100
80
60
40
20
0
UK 1
UK 2
UK 3
US 1
US 2
US 3
(b)
factor omitted
Age 1st symptom
Age at assessment
FEV1
TLCO
post walk saturation
Present dyspnoea
Present pneumothorax
Present angiomyolipoma
ever had lymphatic disease
Ever had pneumothorax
Post-menopause
Surgery for pneumothorax
Present after screening
Has tuberous sclerosis
Ever had angiomyolipoma
Intervention for angiomyolipoma
Full dataset
Posterior prediction
Cluster 1
Cluster 2
Cluster 3
1
.5
1
.5
1
.5


(a)
Model
2 cluster
3 cluster
Cluster
1
2
1
2
3
131
155
207
38
41
ΔFEV1 (ml/yr)
0
-50
-100
-150
129
153
204
37
41
ΔTLCO (mmol/kPa/min/yr)
0.0
-0.1
-0.2
-0.3
-0.4
(b)
Pneumothorax (%)
p=0.0002
Observation (months)
Number at risk
107 97 80 62 41 23 14
109 106 88 74 54 38 28
(c)
Survival (%)
p=0.0045
Observation (months)
Number at risk
130 128 100 11 3
8 8 6 6 1
24 24 13 13 2

Machine learning can predict disease manifestations and outcomes in lymphangioleiomyomatosis.

Chernbumroong S, Johnson JI, Gupta N, Miller S, McCormack FX, Garibaldi J, Johnson SR.

Supplementary methods and results

Supplementary methods

Subjects and clinical data.

The discovery cohort comprised 173 women recruited from the National Centre for LAM in Nottingham UK between 2011 and 2018. A further 10 were added after the discovery analysis until December 2019 and contributed to the outcome analyses. All subjects had LAM defined by current ATS/JRS criteria[1]. All subjects fitting these criteria were invited to participate in the study irrespective of length of follow-up. Outcome analyses included only subjects with follow-up data and the numbers included are described for these analyses individually.

At their first visit to the centre, which formed the baseline assessment for the study, subjects had a clinical assessment, comprising CT of the chest, abdomen and pelvis, screening for TSC, full lung function, bronchodilator reversibility testing and a six minute walk test according to ERS/ATS standards[2]. CT was used to screen for angiomyolipoma at first visit. At follow up visits, clinical outcomes and complications were recorded, FEV_1 and TL_{CO} were repeated and angiomyolipoma size monitored according to a defined protocol at least annually using ultrasound or MRI. Angiomyolipoma causing symptoms, or greater than 4cm in diameter, were discussed with a view to an intervention[3]. All measurements were made as part of clinical care, the study was approved by the East Midlands Research Ethics Committee (13/EM/0264) and participants gave written informed consent.

The replication cohort comprised 186 subjects recruited between 1998 and 2003 to the National Heart Lung and Blood Institute (NHLBI) Registry study on the natural history of LAM[4]. Clinical and serial lung function data were obtained from the National Disease Research Interchange (Philadelphia, USA). All-cause mortality and lung transplantation data for the period until December 2014, were obtained from the United States National Death Index and the United Network for Organ Sharing databases respectively.

In the discovery cohort we collected 25 variables of presumed importance to LAM comprising demographic data (age at presenting symptom, age at assessment and body mass index (BMI)), disease duration (defined by the time from first symptom attributable to LAM to time at baseline assessment), presenting symptom of LAM (one only of dyspnoea, pneumothorax, other respiratory symptom, angiomyolipoma related symptom, other non-respiratory symptom, no LAM related symptoms or diagnosed after screening), clinical phenotype (the presence of any of pneumothorax, angiomyolipoma, TSC or lymphatic manifestations during the whole disease course), oestrogen exposure (menopausal status and the number of children), disease activity (the need for surgical intervention for pneumothorax, intervention for angiomyolipoma and serum VEGF-D level), physiology (percent predicted (%) FEV_1, TL_{CO}, $\%FEV_1/\%TL_{CO}$ at assessment and minimum oxygen saturation during a walk test) and LAM treatment (mTOR inhibitor use, oxygen and transplant referral) were collected for each subject.

The NHLBI cohort collected similar data with some exceptions[4]. Date of diagnosis of LAM, rather than date of first LAM symptom was recorded and minimum oxygen saturation during a walk test, BMI, number of children and serum VEGF-D were either not recorded or unavailable.

Prospective change in lung function for both cohorts was calculated by the regression slope of all FEV_1 (ΔFEV_1) or TL_{CO} (ΔTL_{CO}) values and expressed as change in ml/year or mmol/min/kPa/yr respectively using Excel (Microsoft Corporation). Clinical outcomes including pneumothorax and the need for an intervention for angiomyolipoma was recorded prospectively in the period following baseline assessment.

Serum VEGF-D was determined using Quantikine ELISA DVED00, (R&D Systems, Abingdon, UK).

Machine learning methodology

Data pre-processing

The data set was first pre-processed which involved data cleaning and data validity checking. Three imputation techniques i.e. Multiple Imputation Chain Equation (MICE), Random Forest (RF), and MICE with RF were investigated. An R package MICE[5], and missForest[6] were used in this stage.

To check the imputed data validity, Kolmogorov-Smirnov Tests, Fisher's Test and Pearson's Chi-squared Test were performed to check the distributions between the imputed and original data.

Data transformation

As there are both numerical and categorical variables in the data set, two techniques were used to transform the data so that they are suitable for clustering analysis. The first technique is based on Principal Component Analysis (PCA) with Multiple Correspondent Analysis (MCA). An R package FactoMineR[7] was used to perform this step. The second technique used Gowers distance[8] to calculate the dissimilarity between individuals. Gowers distance compares two variables i and j, at a location k and assign a score $d_{i,j,k}$ of zero if i_k is different from j_k, or a positive value if i_k is similar to j_k. For categorical data, the score d_{ij} is one if values of i_k and j_k is the same, otherwise the score is zero:

$$d_{ij} = d(i,j) = sum(k = 1 : p; w_k \delta(ij;k) d(ij,k)) / sum(k = 1 : p; w_k \delta(ij;k))$$

Optimal number of cluster identification

Seven clustering techniques including K-means, Fuzzy C-means, Partition Around Medoids (PAM), hierarchical clustering, hierarchical + Kmeans, hierarchical + fuzzy, hierarchical + PAM, with varying number of clusters were used to produce cluster models. In this study, number of clusters ranging from 2 to 10 were used. As some of the selected techniques are based on random initialisation, all techniques were repeated 5 times.

A brief definition of techniques used are described below:

Fuzzy C-means (FCM) is a partition-based clustering algorithms. Given the data set X = {x_1, x_2, ..., x_d}, where d is the number of features. The aim of FCM is to minimise the cost function J_m:

$$J_m(U,C) = \sum_{i=1}^{d} \sum_{j=1}^{C} \upsilon_{ij}^{m} \parallel x_i - c_j \parallel^2$$

where c_j is the centre of cluster j, υ_{ij} is the degree of membership of x_i in cluster j and x ∈ [0, 1], $\|*\|$ is the similarity function, and m is the degree of fuzzification and m ∈ [1, ∞). Fuzzy C-mean performs recursively where each iteration the degree of membership υ_{ij}, and the cluster center c_j is updated as follows:

$$\upsilon_{ij} = \frac{1}{\sum_{k=1}^{C} \frac{\|x_i - c_j\|}{\|x_i - c_k\|}^{\frac{2}{m-1}}}$$

$$c_j = \frac{\sum_{i=1}^{d} \upsilon_{ij}^{m} x_i}{\sum_{i=1}^{d} \upsilon_{ij}^{m}}$$

The iteration stops when $\max_{ij} \mid \upsilon_{ij}^{k+1} - \upsilon_{ij}^{k} \mid < \varepsilon$, where ε is the termination criteria and ε ∈ [0, 1].

K-mean is another clustering algorithm. Similar to Fuzzy-cmean, clusters k clusters $\{\square_{\square}\}_1^{\square}$ are chosen which aim to minimise:

$$\sum_{j=1}^{k} \inf_{y_j \in \mathcal{X}} \sum_{i \in C_j} \|x_i - y_j\|^2.$$

Where inf is realisable and is □(□_□). In this paper, Hartigan and Wong[9] algorithm was adopted, which employed greedy heuristic search to repeatedly select point by point, and determine its optimal cluster assignment[10].

Partitioning Around Medoids (PAM)
PAM is another partition-based clustering algorithm. Unlike K-mean which relies on Euclidean geometry to estimate clusters' centers, PAM uses medoid, the object with the smallest dissimilarity to all others in the cluster. This allows complex distance functions e.g. Jaccard, Gower to be used[11]. The aim is to find k representative objects which minimize the sum of the dissimilarities of the observations to their closest representative object.

In this study, Euclidean distance was used to calculate the dissimilarity between individuals, and Ward's method was used for clustering in hierarchical clustering algorithm.

25 internal cluster validity indexes, and Gap statistics[12] with bootstrap were used to measure the cluster validity. In this study the number of bootstrap was set to 100. A majority voting was used to determine the optimal number of clusters for each algorithm. In the case of ties, the smallest number of clusters is selected.

Euclidean distance was used to calculate the dissimilarity between individuals. The Euclidean distance calculates straight-line distance between two points:

$$\| x - y \| = \sqrt{\sum_{i=1}^{n} | x_i - y_i |^2}$$

Wards method was used for clustering in hierarchical clustering algorithm. 25 internal cluster validity indexes, and Gap statistics with bootstraping were used to measure the cluster validity. The list of validation indexes and their criteria are presented in Table 1. We applied all the criteria on cluster results and select the optimal number of clusters based on majority voting with the number of bootstraps set to 100. In the case of ties, the smallest number of clusters is selected.

Cluster analysis
4 algorithms, Kmeans, Fuzzy C-means, PAM, and hierarchical clustering were used for cluster analysis. As each algorithm may be assigned different cluster number (for example, cluster 1 in kmeans may be the same as cluster 2 in PAM), it is necessary to inspect the cluster plots, and reassign clusters before combining results. The results from the four techniques were later combined based on majority voting to give the final results. To identify the prominent characteristics in each cluster, Fisher's Test, Pearson's Chi-squared Test, Wilcoxon Rank Sum and Signed Rank Tests were used at 95% confidence interval.

We identified the smallest number of variables necessary to classify subjects with LAM into the groups defined in the earlier stage. Three feature selection techniques, Recursive Feature Elimination (RFE), Correlation-based feature selection (CFS) and Maximum Relevance Minimum Redundant (MRMR) were used to identify the markers necessary for classifying subjects into clusters. Here we briefly describe algorithms used:

RFE combines a backward search with classification algorithms to identify the optimal subset of features. First, a model is built and evaluated based on all features and a feature ranking performed. In each iteration, the least significant feature is removed, and a model built and evaluated based on the remaining features, with the process is repeated until no features are left. The optimal subset of features was selected based upon the feature subset with the greatest accuracy. Feature importance was calculated using Naive Bayes (NB) and RF

with RFE. For NB, the feature ranking uses a filter method based on the Area Under the Receiver Operating Characteristic curve (AUC). For each variable, AUC is calculated using different cut-off points. Important features are ones with high AUC. For RF, features are ranked based on the mean decrease in node impurities from splitting on the variable. Gini Index (GI) is used as a measurement for the node impurity. Given at any splitting point in a tree, GI of a variable V with n possible values V = {v_1, v_2, ..., v_n}, GI can be calculated as:

$$GI_V = 1 - \sum_{v \in V} P(v)^2$$

where P(v) is the probability of event v after the split.

GI is averaged over all trees. Important features are ones with high mean decrease in GI (i.e. features that make the node become purer).

- CFS is based on correlation between features where it is believed that features are useful if it is correlated with outcomes and uncorrelated with each other. A feature is selected if it predicts outcomes in spaces which have not already predicted by other features. A score of a feature set S with k variables is calculated as:

$$Scos = \frac{k\overline{\tau_{cf}}}{\sqrt{k + k(k-1)\overline{\tau_{ff}}}}$$

where τcf is the mean feature-class correlation, and τf f is the average feature-feature correlation.
A best-first-search strategy was employed with CFS with feature selection carried out using 5-fold cross-validation (CV) with 10 runs, with all results aggregated. The ranking was used in a sequential feature forward selection (SFS) with 5-fold CV repeated for 10 times to obtain the final evaluation.

- MRMR is a feature selection based on mutual information which measures how much information one variable has on another variable. Given two variables with discrete values x = {x_1, x_2, ..., x_i} and y = {y_1, y_2, ..., y_j}, where i, j are the numbers of possible values in x, y. Their mutual information (MI) was calculated as:

$$MI(x;y) = \sum_{i,j} p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(x_i)p(y_j)}$$

To measure class relevancy, mutual information between a variable and outcomes was calculated to identify variables with high discriminant power. Given a variable x, and outcome o = {o_1, ..., o_k}, relevancy was calculated as:

$$Rel_i = \frac{1}{|S|} \sum_{i \in S} MI(x_i; o)$$

Redundancy was reduced within the selected features by selecting a feature with minimum redundancy, with redundancy between two variables x, y can be calculated as:

$$Red_i = \frac{1}{|S|^2} \sum_{i,j \in S} MI(x_i; y_i)$$

MRMR selects a feature set based that maximizes discriminant power between a variable and outcomes, whilst minimising redundancy between variables. Relevancy and redundancy were optimised using Mutual Information Difference criterion max and Mutual Information Quotient criterion

– Mutual Information Difference criterion

$$\max(Rel_i - Red_i)$$

– Mutual Information Quotient criterion

$$\max(\frac{Rel_i}{Red_i})$$

Data were first discretized into three stages where stage -1 represents data value lower than μ − σ/2, stage, 1 represents data value greater than μ + σ/2, and stage 0 represents value between μ − σ/2, μ + σ/2. 25 Similar to CFS, feature ranking was performed using MRMR based on 5-fold CV and 10 runs. Then, the ranking was used in SFS to identify feature set and its performances.

Classification models

Classification models were developed using the cluster results from the previous stage. First, feature selection technique namely Recursive Feature Elimination (RFE) was used to identify the marker necessary for classifying LAM patients. Five classification algorithms i.e. Naïve Bayes (NB), RF, C4.5, C5.0, and CART, were investigated.

CART, C4.5 and C5.0 are variant of decision tree (DT) algorithms[13,14]. DT tries to find the variables that split the data such that the data become as pure as possible. The process of dividing continues until data cannot be split further. Often, a tree pruning process is applied to reduce overfitting. In CART, a binary tree is constructed where each node only contains only two sub-nodes. The three algorithms use different splitting, and pruning criteria. for splitting criteria, CART uses towing criteria, whereas C4.5 uses Gain ratio, C5.0 uses Information Gain. For pruning criteria, CART uses cost-complexity, while C4.5 uses error-based pruning, and C5.0 performs 2 stages pruning i.e. individual branch, and global tree.

RF was developed by Breiman[15] and is based on ensemble learning method. In RF algorithm, □_□□□□ bootstrap samples are drawn from the data, and tree models are developed from them. The final results are the aggregation of individual tree outcomes[16].

NB is a probabilistic classifier based on Bayes theorem and chain rule. Given an event D of a patient having a disease or not such that D ∈ {d^+, d^-}, and test result T ∈ {t^+, t^-}, the posterior probability of a patient having a disease given that the test result is positive, P($d^+|t^+$), was calculated as:

$$P(d^+|t^+) = \frac{P(t^+|d^+)P(d^+)}{P(t^+)}$$

The posterior probability of the event was based upon a chain rule simplified to:

$$P(A, B, C) = P(A|B, C)P(B|C)P(C)$$

Giving the probability of a patient having a disease given that test 1 is positive and test 2 is negative is:

$$\begin{aligned} P(d^+|t_1^+, t_2^-) &= \frac{P(t_1^+, t_2^-|d^+)P(d^+)}{P(t_1^+, t_2^-)} \\ &= \frac{P(t_2^-|d^+)P(t_1^+|t_2^-, d^+)P(d^+)}{P(t_1^+|t_2^-)P(t_2^-)} \\ &= \frac{P(t_2^-|d^+)P(t_1^+|t_2^-)P(d^+)}{P(t_1^+|t_2^-)P(t_2^-)} \\ &= \frac{P(t_2^-|d^+)P(d^+)}{P(t_2^-)} \end{aligned}$$

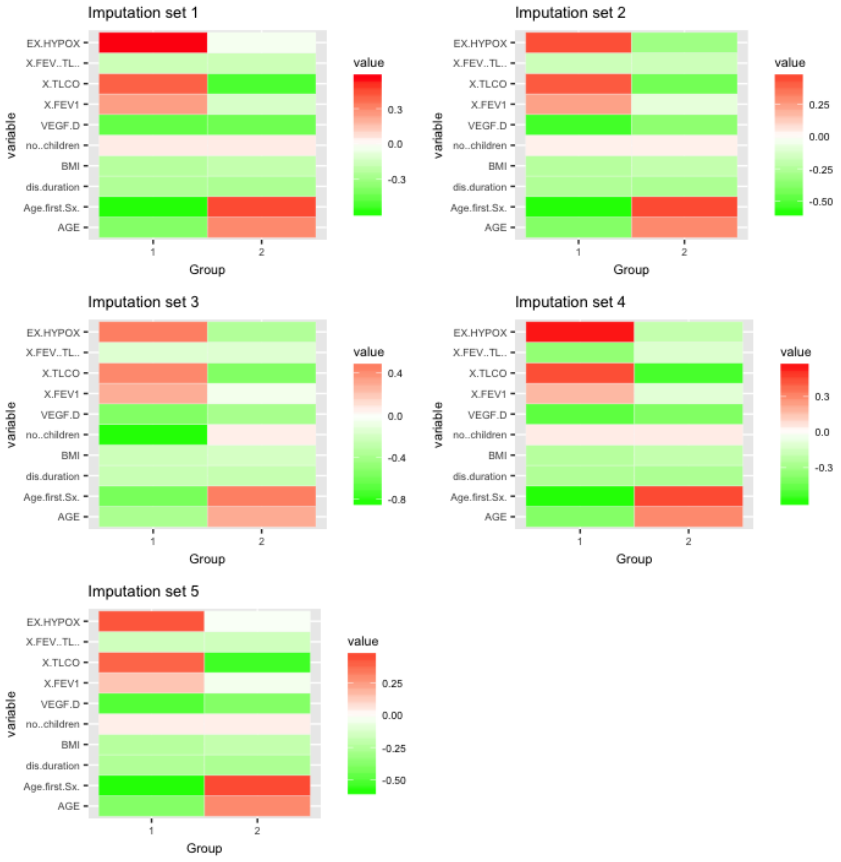
Supplementary Results

Data imputation.
173 patients with definite outcome were included in the analysis. 10 independent variables, and all outcome variables contain missing data

25 independent variables were used in data imputation. The results from the statistical analysis revealed that all three imputation techniques produced data similar to the original data. In this study, we chose to use imputed data produced using MICE.

Supplementary figure S1. Data distribution before and after imputation using multiple imputation by MICE.

The figure shows heat maps of the standardized values for each variable compared between cluster 1 and cluster 2 for the five data imputation sets used. Higher values compared to the average value are represented by increasing red colour, lower by green. All imputed data sets are similar indicating the characteristics of the two clusters are consistent across the imputed data sets. A combination of Kolmogorov-Smirnov Tests, Fisher's Test and Pearson's Chi-squared Test were performed and showed there were no differences in the distributions between the imputed and original data sets.



Defining the optimal number of clusters.
All 5 imputed data sets obtained were used to identify the optimal number of clusters, and the results are shown in Table E1. Based on majority voting, the number of optimal clusters was 2.

Supplementary table S1. Output of optimal cluster number methodology.

Input	Method	Validation technique	Cluster no.
PCA+MCA transformed	Kmean	Internal cluster indexes	2
PCA+MCA transformed	Fuzzy c mean	Internal cluster indexes	8
PCA+MCA transformed	PAM	Internal cluster indexes	10
PCA+MCA transformed	Hierarchy	Internal cluster indexes	2
PCA+MCA transformed	Hierarchy + Kmean	Internal cluster indexes	2
PCA+MCA transformed	Hierarchy + PAM	Internal cluster indexes	10
PCA+MCA transformed	Hierarchy + Fuzzy c mean	Internal cluster indexes	9
Gower distance	PAM	Internal cluster indexes	2
Gower distance	Hierarchy	Internal cluster indexes	3
PCA+MCA transformed	Kmeans	Gap statistics	2
PCA+MCA transformed	PAM	Gap statistics	9
PCA+MCA transformed	Hierarchy + Kmean	Gap statistics	2
PCA+MCA transformed	Fuzzy c mean	Gap statistics	2

Input comprises data transformed by principal component analysis (PCA) and multiple correspondent analysis (MCA) for numerical and categorical variables repectively. PAM, partitioning around medoids.

Predictive modelling
We investigated 5 Machine Learning techniques i.e. Random Forest (RF), Decision Tree, CART, C4.5, C5.0, and Naive Bayes (NB) in data modelling. The experiment was carried out using 5-fold cross validation and repeated 10 times. Supplementary table S2 shows the performances of each technique.

Supplementary table S2. Performance of machine learning techniques.

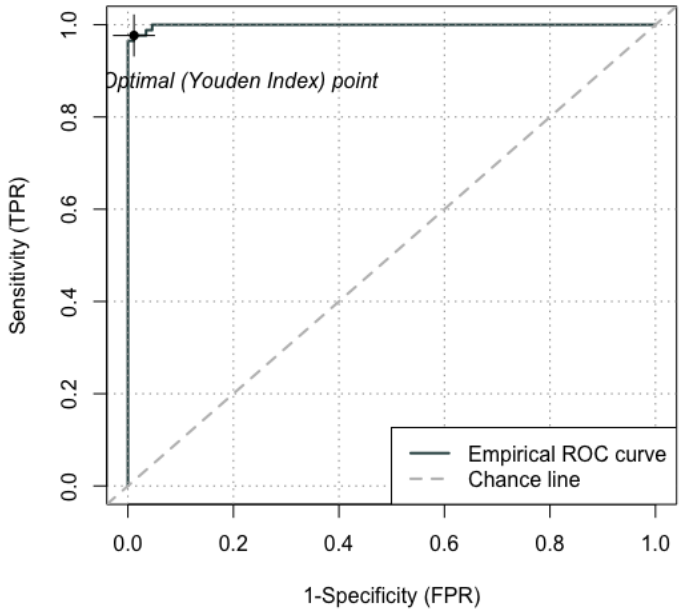
	RF	C4.5	C5.0	CART	NB
AUC	0.99317	0.89889	0.925416	0.868421	0.989501
prAUC	0.991595	0.433186	0.89319	0.782929	0.986296
Accuracy*	0.958	0.884824	0.881765	0.833882	0.960588
Kappa	0.915985	0.769672	0.763559	0.667776	0.921145
F1	0.958548	0.884973	0.881817	0.83462	0.961459
Sensitivity	0.96	0.875814	0.872093	0.828605	0.97186
Specificity	0.955952	0.894048	0.891667	0.839286	0.949048
Pos Pred Value	0.957126	0.894435	0.891892	0.840738	0.951289
Neg Pred Value	0.958952	0.875568	0.872065	0.827089	0.970559
Precision	0.957126	0.894435	0.891892	0.840738	0.951289
Recall	0.96	0.875814	0.872093	0.828605	0.97186
Detection Rate	0.485647	0.443059	0.441176	0.419176	0.491647
Balanced Accuracy	0.957976	0.884931	0.88188	0.833945	0.960454

Accuracy is the ability of the model to correctly assign patients to cluster 1 or 2. *since the number of patients in each cluster is inequivalent the balanced accuracy i.e. balanced using the proportion of patients is a better estimation of the accuracy. Sensitivity, the ability to correctly identify those patients with cluster 1 characteristics. Specificity, the ability to correctly identify patients with cluster 2 characteristics. PPV (positive predictive value) quantifies the likleyhood that a patient has the characteristics of cluster 1 given a positive result (predicting C1). The NPV (negative predictive value) quantifies the likleyhood that a patient has the characteristics of cluster 2 given a negative result (predicting C2). P-Value is the accuracy over the no Information Rate (probability of correctly identifying patient with cluster 1 characteristics without given variable data).

Naive Bayes Model

The performance of the Naive Bayes Model for the two-cluster model is shown below.

Supplementary figure S2. Receiver operating characteristic curve for the Naive Bayes Model for the two-cluster model

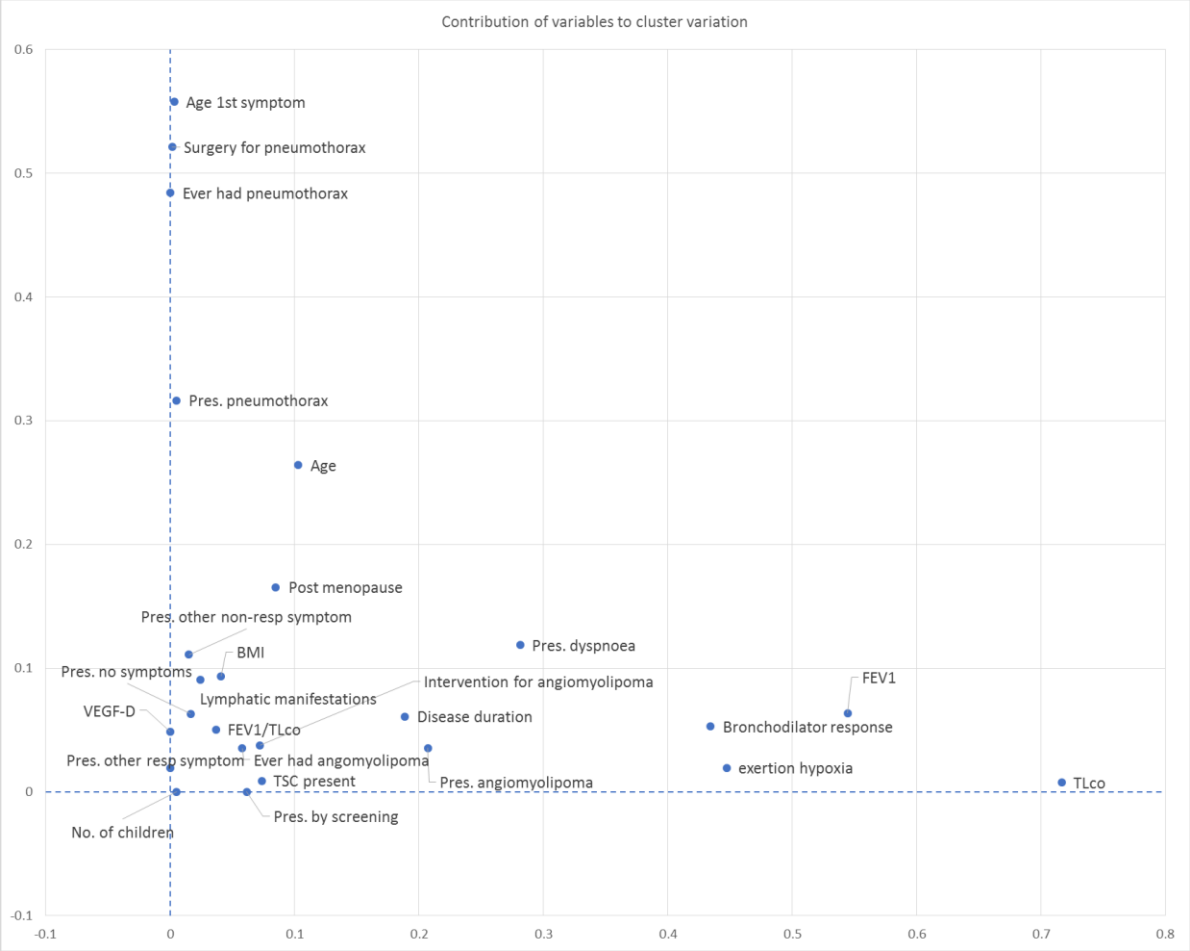


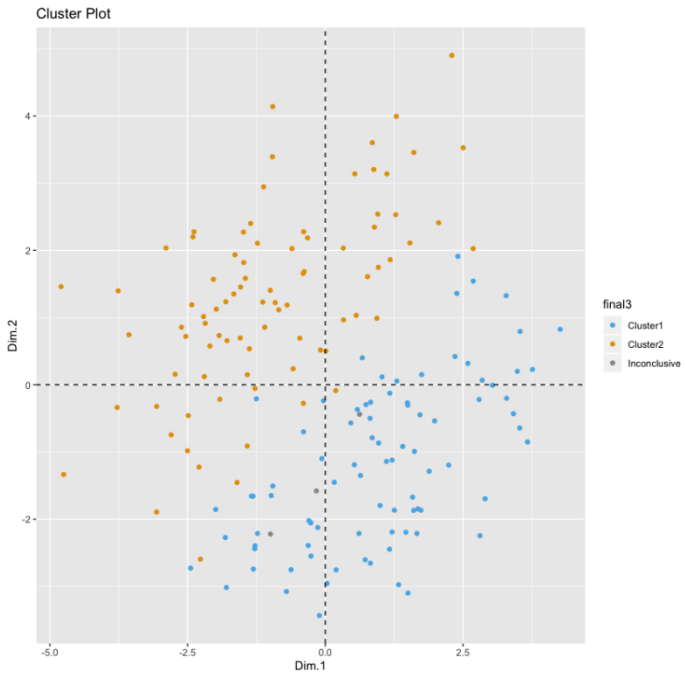
Confusion Matrix and Statistics for the two-cluster model
(Positive' Class: G1)

Prediction G1 vs G2
Accuracy : 0.9827 (95% CI: 0.9502 - 0.9964)
No Information Rate (NIR): 0.5029
P-Value [Acc > NIR] : <2e-16
Kappa : 0.9653
Mcnemar's Test P-Value : 0.2482
Sensitivity : 1.0000
Specificity : 0.9651
Pos Pred Value : 0.9667
Neg Pred Value : 1.0000
Prevalence : 0.5029
Detection Rate : 0.5029
Detection Prevalence : 0.5202
Balanced Accuracy : 0.9826

Supplementary figure S3 Contribution of variables to cluster variation.

The following figures show variables contribute to each cluster. For example, patients on the top of the graph (Cluster 2) are older and were older at first symptom, compared with patients in the bottom of the graph (Cluster 1). Patients on the right side of the graph (Cluster 2) have higher TL_{CO} with those on the left side of the graph (Cluster 1).





Dot plot showing separation of individuals into clusters by principal components.

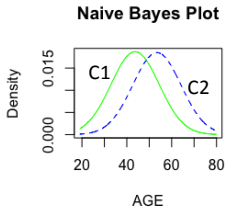
Variance of model factors

Data used for training are imputed data from MICE and removed duplicates (N=427). Labels are from combining multiple clusters

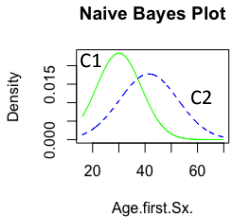
Supplementary table S3. The following graphs and table show the probability of a patient falling into each cluster dependent on each factor. For example, a patient presenting with shortness of breath is more likely to be in cluster 1. Older patients are likely to be in cluster 2.

Two cluster model

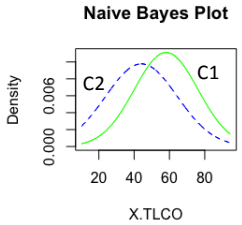
AGE



Age at first Sx.



TLCO



Presentation with shortness of breath

	No	Yes
Cluster1	0.91954023	0.08045977
Cluster2	0.29069767	0.70930233

Presentation with pneumothorax

	No	Yes
Cluster1	0.47126437	0.52873563
Cluster2	0.98837209	0.01162791

Presentation with angiomyolipoma

	No	Yes
Cluster1	0.71264368	0.28735632
Cluster2	0.97674419	0.02325581

Presentation with no symptoms

	No	Yes
Cluster1	1.00000000	0.00000000
Cluster2	0.91860465	0.08139535

Ever had angiomyolipoma

	No	Yes
Cluster1	0.2643678	0.7356322
Cluster2	0.4651163	0.5348837

Ever had lymphatic manifestations

	No	Yes
Cluster1	0.94252874	0.05747126
Cluster2	0.72093023	0.27906977

Ever had pneumothorax

	No	Yes
Cluster1	0.2758621	0.7241379
Cluster2	0.8488372	0.1511628

Post Menopause

	No	Yes
Cluster1	0.1954023	0.8045977
Cluster2	0.4767442	0.5232558

Surgery for pneumothorax

	No	Yes
Cluster1	0.3678161	0.6321839
Cluster2	0.8837209	0.1162791

Intervention for angiomyolipoma

	No	Yes
Cluster1	0.5862069	0.4137931
Cluster2	0.8255814	0.1744186

Supplementary table S4. Statistical tests for numerical variables in two-cluster model. Continuous variables are analysed in section 9a), categorical in section (b) and the differences between variables for each group are presented in (c).

(a)

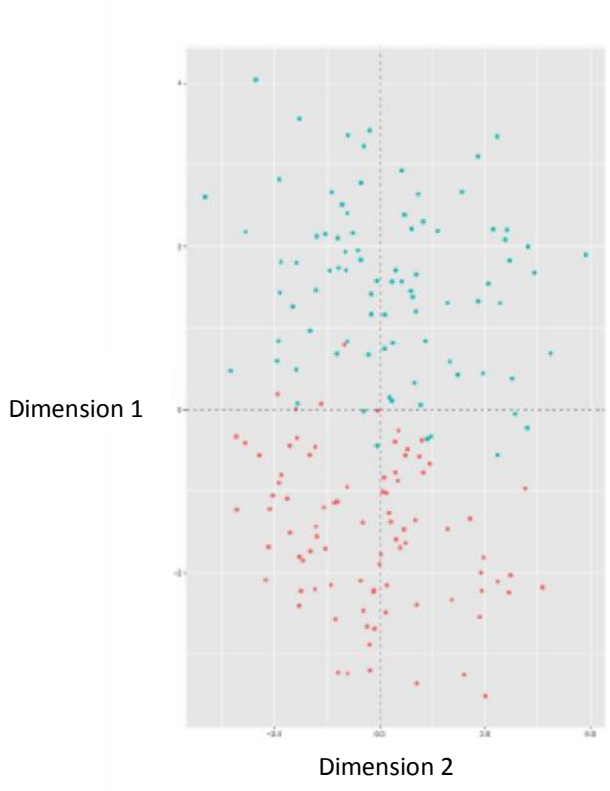
Continuous variable	Statistics	Direction	Cluster average	Global average	P value
Cluster 1					
AGE	5759.5	-	44.0	48.0	0.002016
Age first symptom	5320.5	-	29.0	34.5	0.000116
TLCO	9050.0	+	59.3	51.6	0.007728
Cluster 2					
AGE	9205.0	+	52.00	48.0	0.001858
Age first symptom	9644.0	+	40.90	34.5	0.0001032
TLCO	5914.5	-	40.55	51.6	0.007267

(b)

Categorical variable	Cluster average	Statistics	Per model cluster	Per model global	Global average	P value
Cluster 1						
present dyspnoea	No	71.688492	91.95402	76.19048	No	0.0004998
present pneumothorax	Yes	NA	52.87356	97.87234	No	2.44e-16
present angiomyolipoma	Yes	NA	71.26437	42.46575	No	9.995e-07
present no symptoms	No	NA	100.0000	52.40964	No	0.006603
ever had angiomyolipoma	Yes	7.527227	73.56322	58.18182	Yes	0.007996
ever had lymphatic disease	No	15.220782	94.25287	56.94444	No	0.0009995
pneumothorax	Yes	57.643460	72.41379	82.89474	No	0.0004998
post menopause	Pre	15.360550	80.45977	60.86957	Pre	0.0009995
surgery for pneumothorax	Yes	49.075631	63.21839	84.61538	No	0.0004998
intervention for angiomyolipoma	Yes	11.920365	58.62069	41.80328	No	0.001499
Cluster 2						
present dyspnoea	Yes	71.688492	70.93023	89.70588	No	0.0004998
present pneumothorax	No	NA	98.83721	67.46032	No	2.44e-16
present angiomyolipoma	No	NA	97.67442	57.53425	No	9.995e-07
present no symptoms	Yes	NA	91.86047	47.59036	No	0.006603
ever had angiomyolipoma	No	7.527227	53.48837	41.81818	Yes	0.004998
ever had lymphatic disease	Yes	15.220782	72.09302	43.05556	No	0.0004998
pneumothorax	No	57.643460	84.88372	75.25773	No	0.0004998
post menopause	Post	15.360550	52.32558	39.13043	Pre	0.0004998
surgery for pneumothorax	No	49.075631	88.37209	70.37037	No	0.0004998
intervention for angiomyolipoma	No	11.920365	82.55814	58.19672	No	0.001999

(c)

Variables	Cluster 1	Cluster 2
AGE	-	+
Age first symptom	-	+
TLCO	+	-
present dyspnoea	No	Yes
present pneumothorax	Yes	No
present angiomyolipoma	Yes	No
present no symptoms	No	Yes
ever had angiomyolipoma	Yes	No
ever had lymphatic disease	No	Yes
pneumothorax	Yes	No
post menopause	Pre	Post
surgery for pneumothorax	Yes	No
intervention for angiomyolipoma	Yes	No

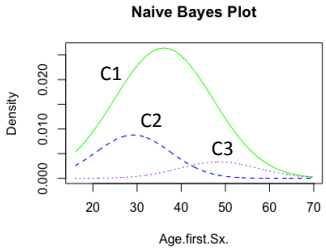


Supplementary Figure S4. Subject distribution in the two-cluster model. Cluster 1 subjects shown in blue, cluster 2 red.

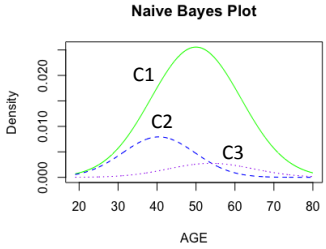
Three-cluster model

Supplementary table S5. The following graphs and table show the probability of a patient falling into each cluster according to each factor. For example, a patient with angiomyolipoma is more likely to be in cluster two.

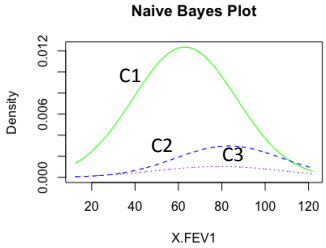
Age at first symptom



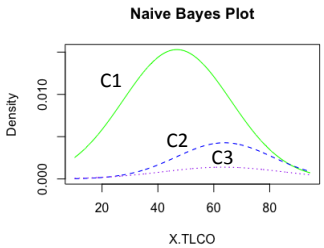
Age at presentation



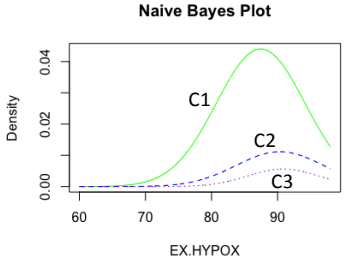
FEV_1



TL_{CO}



exertional hypoxaemia



Presentation with dyspnoea

	No	Yes
Cluster 1	0.484375	0.515625
Cluster 2	0.937500	0.062500
Cluster 3	1.000000	0.000000

Presentation with pneumothorax

	No	Yes
Cluster 1	0.6328125	0.3671875
Cluster 2	1.0000000	0.0000000
Cluster 3	1.0000000	0.0000000

Presentation with angiomyolipoma

	No	Yes
Cluster 1	1.00000	0.00000
Cluster 2	0.15625	0.84375
Cluster 3	1.00000	0.00000

Ever had lymphatic manifestations

	No	Yes
Cluster 1	0.8515625	0.1484375
Cluster 2	0.8750000	0.1250000
Cluster 3	0.5384615	0.4615385

Ever had pneumothorax

	No	Yes
Cluster 1	0.48437500	0.51562500
Cluster 2	0.71875000	0.28125000
Cluster 3	0.92307692	0.07692308

Post menopause

	No	Yes
Cluster 1	0.3828125	0.6171875
Cluster 2	0.0937500	0.9062500
Cluster 3	0.4615385	0.5384615

Surgery for pneumothorax

	No	Yes
Cluster 1	0.57812500	0.42187500
Cluster 2	0.68750000	0.31250000
Cluster 3	0.92307692	0.07692308

Presentation after screening

	No	Yes
Cluster 1	1.00000	0.00000
Cluster 2	0.84375	0.15625
Cluster 3	1.00000	0.00000

TSC present

	No	Yes
Cluster 1	0.8515625	0.1484375
Cluster 2	0.6250000	0.3750000
Cluster 3	0.6153846	0.3846154

Angiomyolipoma present

	No	Yes
Cluster 1	0.4296875	0.5703125
Cluster 2	0.0000000	1.0000000
Cluster 3	0.6153846	0.3846154

Intervention for angiomyolipoma

	No	Yes
Cluster 1	0.78125000	0.21875000
Cluster 2	0.31250000	0.68750000
Cluster 3	0.92307692	0.07692308

Supplementary table S6. Statistical tests for numerical variables in three-cluster model. Continuous variables are analysed in section (a), categorical in section (b) and the differences between variables for each group are presented in (c).

(a)

Continuous variable	Statistics	Direction	Cluster average	Global average	P value
Cluster 1					
Age first symptom	1822	-	28.25	34.5	0.00216
Age	1648	-	38.00	48.0	0.00028
FEV_1	3688	+	90.90	68.8	0.00286
TL_{CO}	3747	+	64.55	51.6	0.00150
Cluster 3					
Age first symptom	1829.5	+	46.8	34.5	0.00017

(b)

Categorical variable	Cluster average	Statistics	Per model cluster	Per model global	Global average	P value
Cluster 1						
present dyspnoea	Yes	NA	51.56250	97.05882	No	2.787e-09
present pneumothorax	Yes	NA	63.28125	64.28571	No	8.131e-08
present angiomyolipoma	No	NA	100.0000	87.67123	No	5.945e-20
ever had pneumothorax	Yes	11.6362	51.56250	86.84211	No	0.0009995
post menopause	Post	4.99292	61.71875	68.69565	Pre	0.03098
surgery for pneumothorax	Yes	4.46877	57.81250	68.51852	No	0.04948
present with screening	No	NA	100.0000	76.19048	No	0.001003
TSC	No	10.6269	85.15625	79.56204	No	0.0004998
ever had angiomyolipoma	No	9.12484	57.03125	66.36364	Yes	0.001999
intervention for angiomyolipoma	No	13.6892	78.12500	81.96721	No	0.0004998
Cluster 2						
present dyspnoea	No	NA	93.750	28.57143	No	6.361e-06
present pneumothorax	No	NA	100.000	25.39683	No	1.417e-05
present angiomyolipoma	Yes	NA	84.375	100.0000	No	6.977e-27
ever had pneumothorax	No	3.98206	71.875	23.71134	No	0.04298
post menopause	Pre	NA	90.625	25.21739	Pre	0.0008591
surgery for pneumothorax	Yes	NA	84.375	16.07143	No	0.0001653
present with screening	Yes	6.637398	62.500	14.59854	No	0.01599
TSC	Yes	NA	100.000	29.09091	Yes	8.33e-08
ever had angiomyolipoma	Yes	29.1250	68.750	43.13725	No	0.0004998
Cluster 3						
present dyspnoea	Yes	NA	51.56250	97.05882	No	2.787e-09
present pneumothorax	Yes	NA	63.28125	64.28571	No	8.131e-08

present angiomyolipoma	No	NA	100.0000	87.67123	No	5.945e-20
ever had pneumothorax	Yes	11.63621	51.56250	86.84211	No	0.0009995
post menopause	Post	4.992919	61.71875	68.69565	Pre	0.03098
surgery for pneumothorax	Yes	4.468773	57.81250	68.51852	No	0.04948
present with screening	No	NA	100.0000	76.19048	No	0.001003
TSC	No	10.62689	85.15625	79.56204	No	0.0004998
ever had angiomyolipoma	No	9.124839	57.03125	66.36364	Yes	0.001999
intervention for angiomyolipoma	No	13.68919	78.12500	81.96721	No	0.0004998

(C)

Variables	**Cluster 1**	**Cluster 2**	**Cluster 3**
Age first symptom	-	+	
Age	-		
FEV_1	+		
TL_{CO}	+		
present dyspnoea	Yes	No	No
present pneumothorax	Yes	No	No
present angiomyolipoma	No	Yes	
ever had pneumothorax	Yes	No	No
post menopause	Post	Pre	
surgery for pneumothorax	Yes		No
present with screening	No	Yes	
TSC	No	Yes	
ever had angiomyolipoma	No	Yes	
intervention for angiomyolipoma	No	Yes	
ever had lymphatic disease			Yes

Supplementary Table S7. Two-cluster model applied to the NHLBI cohort.

Factor	Cluster 1	Cluster 2	Mean diff	p
n	82	99		
Demographic *				
Age at assessment (yrs)	41.3 (8.4)	47.9 (8.9)	-6.7	$6.3x10^{-7}$
Age 1st symptom (yrs)	34.1 (8.4)	41.5 (10.5)	-7.4	$4.1x10^{-7}$
Presenting symptom †				
Dyspnoea	1.2	50.5	-38.5	< .00001
Pneumothorax	69.9	5.1	64.8	< .00001
Other respiratory	12.0	52.5	-40.5	< .00001
Angiomyolipoma	9.6	0	9.6	.0053
Screened	7.2	0	7.2	.03
Chance finding	0	39.4	-39.4	< .00001
Phenotype †				
Ever had pneumothorax	79.5	12.1	67.4	< .00001
Ever had angiomyolipoma	22.9	8.1	14.8	< .00001
Lymphatic disease	10.8	21.2	-10.4	.054
TSC	15.7	5.1	10.6	.0002
Lung function *				
FEV_1 (% predicted)	78.7 (25.5)	72.1 (24.6)	6.6	0.068
DL_{CO} (% predicted)	63.5 (22.9)	52.7 (20.6)	10.7	0.00095
Bronchodilator reversibility (%)	11.3 (10.3)	11.9 (10.5)	-0.6	0.67

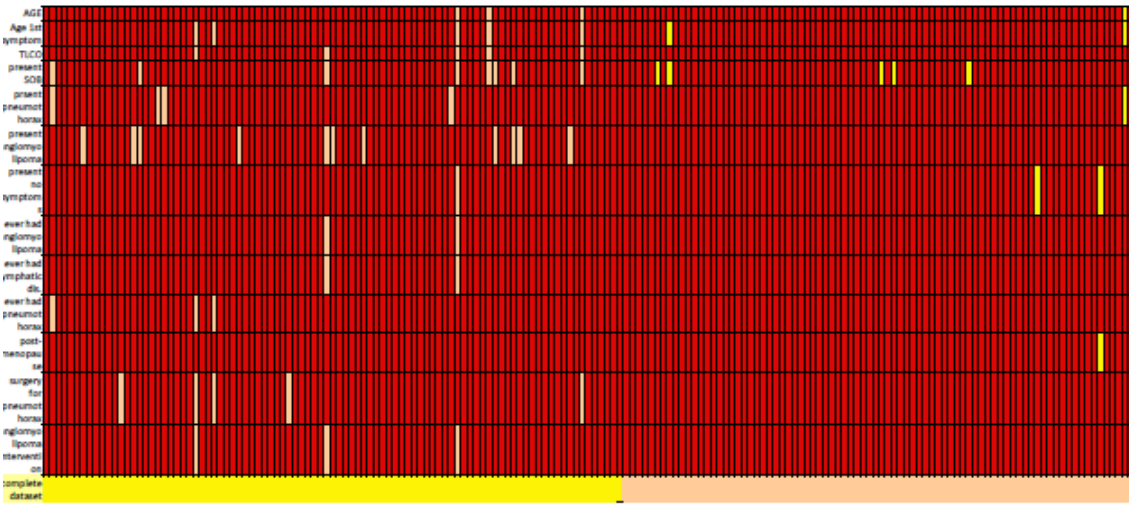
* Mean value (standard deviation) compared by unpaired 2 tail t-test. † Percentage of cohort with this feature present compared by chi square test.

Supplementary Table S8. Three-cluster model applied to the NHLBI cohort.

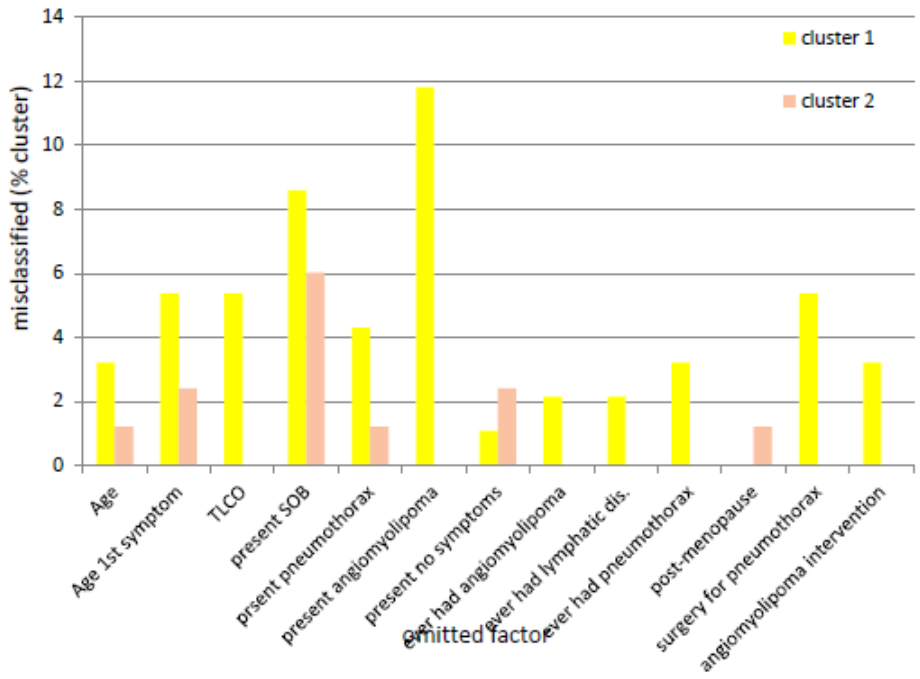
	Cluster 1	Cluster 2	Cluster 3	p
Number in cluster	145	11	28	
Demographic *				
Age at assessment (yrs)	44.7 (9.4)	39.4 (9.2)	43.3 (7.4)	.026
Age 1st symptom (yrs)	37.5 (10.7)	34.4 (6.7)	43.3 (7.4)	.040
Presenting symptom †				
Dyspnoea	35.9	0	0	< .00001
Pneumothorax	43.4	0	0	< .00001
Other respiratory	35.9	0	39.2	< .00001
Angiomyolipoma	0	72.7	0	< .00001
Screened	0.7	27.3	7.1	< .00001
Chance finding	20.7	0	35.7	< .00001
Phenotype †				
Ever had pneumothorax	53.1	9.1	0	< .00001
Ever had angiomyolipoma	8.2	100	14.3	< .00001
Lymphatic disease	12.4	0	42.8	< .00001
TSC	5.5	35.4	21.4	< .00001
Lung function *				
FEV_1 (% predicted)	70.4 (24.1)	91.2 (21.7)	92.3 (21.7)	<.0001
DL_{CO} (% predicted)	53.3 (20.6)	75.5 (23.1)	72.1 (20.7)	<.0001
Bronchodilator reversibility (%)	12.7 (10.6)	8.6 (8.2)	7.5 (9.1)	.037

* mean (+/-SD), analysed by one way ANOVA. † percentage of cohort, analysed by chi square test.

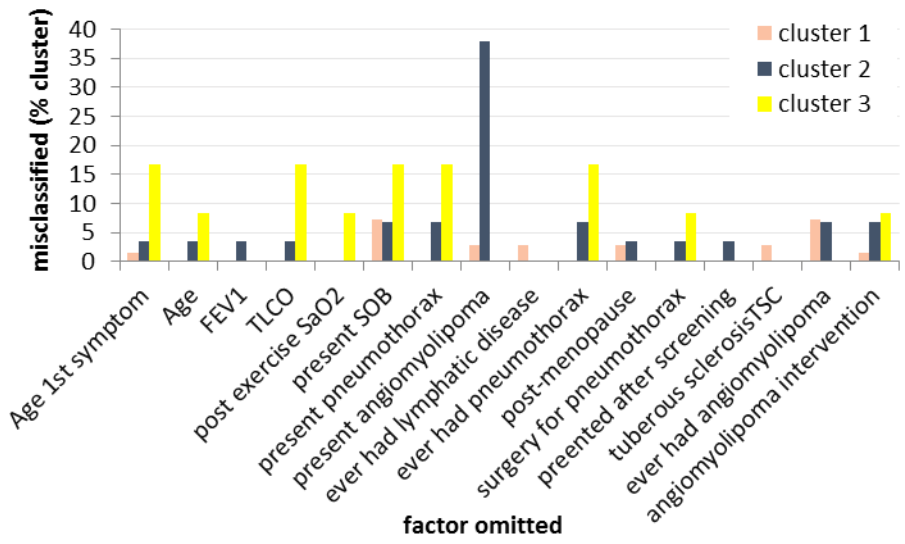
(A)



(B)



Supplementary Figure S5 (a) Effect of missing data upon cluster assignment in the two-cluster model. 112 subjects from the UK cohort with complete data were assigned to clusters and then reassigned with each variable removed in turn. The heatmap is red for correctly assigned subjects (columns) and tan when omission of that variable listed (rows) led to mis-assignment to cluster and yellow to cluster 2. (b) Percentage of misclassified subjects in each cluster after removal of single variables for each cluster.



Supplementary Figure S6 Effect of missing data upon cluster assignment in the three-cluster model. 112 subjects from the UK cohort with complete data were assigned to clusters and then reassigned with each variable removed in turn. Percentage of misclassified subjects in each cluster after removal of single variables for each cluster.

Supplementary table S9. Prospective lung function change the two-cluster model

Cluster	**1**		**2**		**Difference**	**p**
UK Untreated	mean (SD)	n	mean (SD)	n	(95% C.I.)	
FEV_1 change (ml/yr)	-83 (280)	58	-60 (180)	54	23 (-65 to 112)	0.6
TL_{CO} change (mmol/min/kPa/yr)	-0.21 (.54)	54	-0.18 (.45)	51	0.032 (-.164 to .229)	0.7
observation period (months)	54 (36)		49 (36)			
number of observations	6.1 (3.3)		5.6 (3.5)			
UK Rapamycin treated						
FEV_1 change (ml/yr)	-9 (22)	37	1 (59)	44	10 (-60 to 80)	0.7
TL_{CO} change (mmol/min/kPa/yr)	-0.04 (.35)	37	-0.08 (.17)	44	0.04 (-.158 to .088)	0.5
observation period (months)	42 (22)		451 (27)			
number of observations	6.6 (3.9)		10 (6.3)			
NHLBI Untreated						
FEV_1 change (ml/yr)	-74.7 (115)	73	-103.4 (101)	101	-29 (-60 to 3)	.08
TL_{CO} change (mmol/min/kPa/yr)	-0.25 (.31)	71	-0.22 (.36)	100	0.02 (-0.073 to 0.13)	.56
observation period (months)	38.4 (17)		35.9 (16)			
number of observations	3.8 (1.3)		3.8 (1.3)			

p value obtained by unpaired t-test.

Supplementary Table S10. Prospective lung function change the three-cluster model

Cluster	**1**		**2**		**3**		**p**
UK Untreated	mean (SD)	n	mean (SD)	n	mean (SD)	n	
FEV_1 change (ml/yr)	-76 (252)	71	-51 (213)	27	-88 (191)	14	0.86
TL_{CO} change (mmol/min/kPa/yr)	-0.22 (.47)	70	-0.07 (.62)	27	-0.26 (.54)	14	0.38
observation period (months)	55.1 (39)		43.1 (29)		47.9 (32)		
number of observations	6.3 (3.6)		5.0 (3.0)		5.1 (2.8)		
UK Rapamycin treated							
FEV_1 change (ml/yr)	13 (102)	61	-17 (140)	17	-21 (74)	3	0.54
TL_{CO} change (mmol/min/kPa/yr)	-0.034 (.22)	61	-0.13 (.41)	17	-0.16 (.39)	3	0.38
observation period (months)	52.2 (26)		37.4 (22)		23.2 (8)		
number of observations	9.3 (5.8)		5.9 (3.8)		4.3 (1.2)		
NHLBI Untreated							
FEV_1 change (ml/yr)	-94.9 (107)	136	-17.7 (142)	11	-103 (77)	27	.055
TL_{CO} change (mmol/min/kPa/yr)	-0.22 (.31)	134	-0.27 (.30)	10	-0.26 (.46)	27	.078
observation period (months)	37.2 (17)		39.2 (11)		34.8 (15)		
number of observations	3.8 (1.3)		3.9 (1.1)		3.7 (1.3)		

p value obtained by one way ANOVA.

Supplementary table S11. Prospective UK angiomyolipoma outcome according to cluster models.

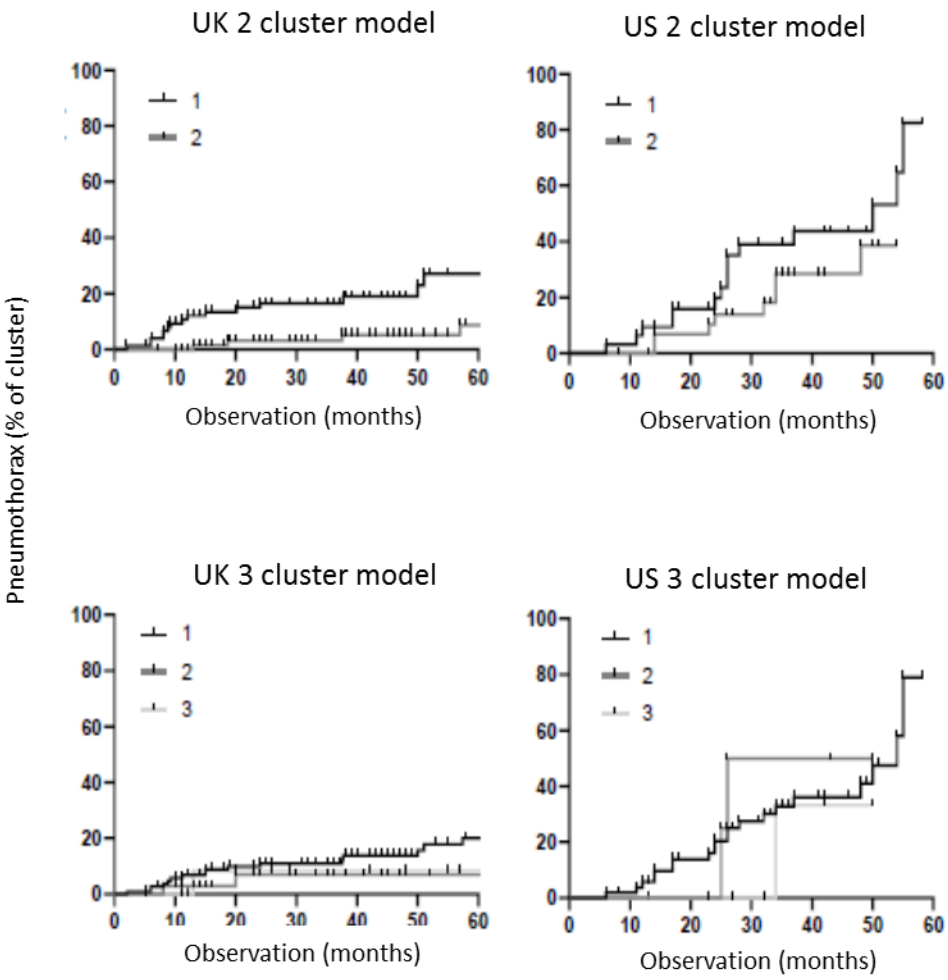
Present = number (percent of cluster) with an angiomyolipoma. int. = number (percent of cluster) requiring an intervention for angiomyolipoma. F/U = mean follow up duration of subjects within that cluster in months. risk = rate of angiomyolpioma intervention / year of follow up of those in cluster with an angiomyolipoma. P = difference in risk between clusters analysed by chi square test.

Cluster	**1**				**2**				**3**				**p**
Model	**present**	**int.**	**F/U**	**risk**	**present**	**int.**	**F/U**	**risk**	**present**	**int.**	**F/U**	**risk**	
2 cluster	69 (72)	16 (17)	47	0.059	43 (50)	5 (6)	52	0.025					<.00001
3 cluster	63 (50)	20 (16)	56	0.069	38 (97)	1 (3)	34	0.001	11 (65)	1 (6)	36	.03	<.00001

Supplementary table S12. Survival outcomes for two and three-cluster models in the NHLBI cohort.

	median time to death or transplant (months)			**hazard ratio** (95% C.I.)			**p**
cluster	**1**	**2**	**3**	**1**	**2**	**3**	
2 cluster	291	285		1.1 (0.66 to 2.0)	0.88 (0.51 to 1.5)		0.907
3 cluster	285	N/A	N/A	N/A	N/A	N/A	0.0045

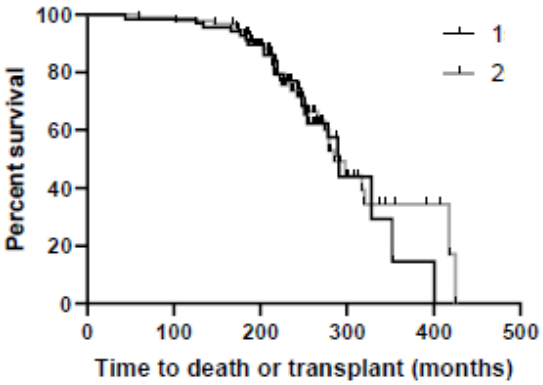
N/A not available: survival >50% at end of analysis.



Supplementary Figure S7. Prospective incidence of pneumothorax in the 2 and 3 cluster models for the UK and NHLBI cohorts. Kaplan Meier analysis of the prospective risk of pneumothorax following cluster assignment in the UK and NHLBI cohorts separately for the two and three cluster model.

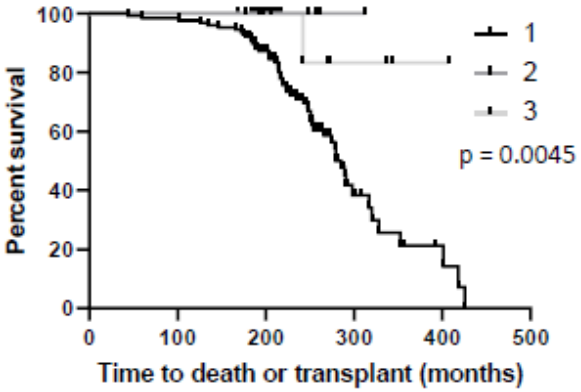
Supplementary Figure S8. Time to death or lung transplant for the 2 and 3 cluster models in the NHLBI cohort. Kaplan Meier analysis of the combined risk of death or need for lung transplantation in the NHLBI cohort stratified using the three cluster model.

2 cluster model



median survival (months)		hazard ratio (95% C.I.)		P
1	2	1	2	
291	285	1.1 (0.66 - 2.0)	0.88 (0.51 - 1.5)	.97

3 cluster model



Supplementary references

1. McCormack FX, Gupta N, Finlay GR, et al. Official American Thoracic Society/Japanese Respiratory Society Clinical Practice Guidelines: Lymphangioleiomyomatosis Diagnosis and Management. American Journal of Respiratory and Critical Care Medicine 2016;**194**(6):748-61.
2. Miller MR, Crapo R, Hankinson J, et al. General considerations for lung function testing. European Respiratory Journal 2005;**26**(1):153.
3. Bee J, Bhatt R, McCafferty I, et al. Audit, research and guideline update: A 4-year prospective evaluation of protocols to improve clinical outcomes for patients with lymphangioleiomyomatosis in a national clinical centre. Thorax 2015.
4. Ryu JH, Moss J, Beck GJ, et al. The NHLBI lymphangioleiomyomatosis registry: characteristics of 230 patients at enrollment. Am J Respir Crit Care Med 2006;**173**.
5. Buuren, S.V. and Groothuis-Oudshoorn, K., 2010. mice: Multivariate imputation by chained equations in R. Journal of statistical software, pp.1-68.
6. Stekhoven, D.J. and Buehlmann, P. (2012), 'MissForest - nonparametric missing value imputation for mixed-type data', Bioinformatics, 28(1) 2012, 112-118, doi: 10.1093/bioinformatics/btr597
7. Le, S., Josse, J. & Husson, F. (2008). FactoMineR: An R Package for Multivariate Analysis. Journal of Statistical Software. 25(1). pp. 1-18
8. Gower, J. C. (1971), A general coefficient of similarity and some of its properties. Biometrics, 27, 623--637.
9. Hartigan, J. A. and Wong, M. A. (1979). Algorithm AS 136: A K-means clustering algorithm. Applied Statistics, 28, 100–108. doi: 10.2307/2346830.
10. Telgarsky, M. and Vattani, A., 2010, March. Hartigan's method: k-means clustering without voronoi. In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (pp. 820-827).
11. Schubert, E. and Rousseeuw, P.J., 2018. Faster k-Medoids Clustering: Improving the PAM, CLARA, and CLARANS Algorithms. arXiv preprint arXiv:1810.05691.
12. Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters 4 in a data set via the gap statistic, Journal of Royal Statistical Society, Series B, 63: 373 511-528.
13 [11] Quinlan, J.R., 2014. C4. 5: programs for machine learning. Elsevier.
14. Lewis, R.J., 2000, May. An introduction to classification and regression tree (CART) analysis. In Annual meeting of the society for academic emergency medicine in San Francisco, California (Vol. 14).
15. Breiman, L., 2001. Random forests. Machine learning, 45(1), pp.5-32.
16. Liaw, A. and Wiener, M., 2002. Classification and regression by randomForest. R news, 2(3), pp.18-22.