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: 2971. Abstract: 245.

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). Over 12 years of follow-up 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 is 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.

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 models to 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.
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.
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.
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.

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.
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.