Unsupervised Phenotypic Clustering for Determining Clinical Status in Children with Cystic Fibrosis

Filipow N[1,2], Davies G[1,5], Main E[1], Sebire NJ[1,5], Wallis C[5], Ratjen F[2,3,4], Stanojevic S[2,6]

ONLINE SUPPLEMENT

Affiliations

[1]UCL Great Ormond Street Institute of Child Health, London UK
[2]Translational Medicine, SickKids Research Institute, Toronto Canada
[3]Division of Respiratory Medicine, Department of Paediatrics, the Hospital for Sick Children, Toronto Canada
[4]University of Toronto, Toronto Canada
[5]Great Ormond Street Hospital for Children and GOSH NIHR BRC, London UK
[6]Department of Community Health and Epidemiology, Dalhousie University, Halifax Canada

Correspondence

Sanja Stanojevic
Department of Community Health and Epidemiology
Dalhousie University
sanja.stanojevic@dal.ca

Take Home Message

Machine learning-derived clusters can be used to define clinical status in children with cystic fibrosis

Keywords

cluster analysis, cystic fibrosis, paediatrics

Acknowledgements

G Davies was supported by a grant from the UCL's Wellcome Institutional Strategic Support Fund 3 [Grant Reference 204841/Z/16/Z]. S Stanojevic received funding from the Program for Individualized Cystic Fibrosis (CF) Therapy Synergy Grant and the European Respiratory Society. N Filipow received funding from a UCL, GOSH and Toronto SickKids studentship. All research at Great Ormond Street Hospital NHS Foundation Trust and UCL Great Ormond Street Institute of Child Health is made possible by the NIHR Great Ormond Street Hospital Biomedical Research Centre. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

Variable Selection for Cluster Model

Assessment of the initially selected 25 Toronto CF (TCF) variables resulted in the exclusion of age at diagnosis, ivacaftor, ethnicity, sex, functional class, pancreatic insufficiency (PI), CF-related diabetes (CFRD) and inhaled antibiotics. Age at diagnosis is less relevant to the CF population since the introduction of new-born screening. There was minimal data for children on ivacaftor and therefore was not a representative descriptor of the population. Ethnicity, sex, functional class, and PI are difficult to coerce into continuous variables and are largely time independent so would provide minimal information on the transition between clusters over time. Functional class and PI are also heavily dominated by classes I-III (94%) and pancreatic insufficiency (92%) and would therefore provide minimal information on variation in the population for defining clusters. Furthermore, the goal of the analysis was to describe all children with CF and to not exclude those without a defined functional class for their mutation. CFRD and inhaled antibiotics were additionally excluded as a result of their categorical nature and were used to corroborate the disease severities of each cluster since they both represent the development of severe disease.

Weight was excluded due to a strong association with body mass index (BMI) ($r = 0.83$). Pulmonary exacerbation (PEx) treated with IV antibiotics in prior year were excluded over hospitalisations in prior year ($r = 0.8$) since hospitalisations encompass most PEx events as well as additional complications. Height and BMI were not strongly correlated ($r = 0.3$), and therefore neither was excluded. In the PCA, the first two principal components combined to explain 24.4% of the variance; the variables with the smallest component loadings which were therefore excluded were deprivation, and rates of previous infection with *Achromobacter sp.*, Methicillin

Resistant *S. aureus* (MRSA) and *B. cepacia* complex. The specific microbiology exclusions were further confirmed by the research team, since very few visits (< 3%) had positive cultures.

The variable exclusions resulted in 11 variables available for iterative clustering: BMI, height, PEx treated with oral antibiotics, hospitalisations in prior year, cough, age, and previous rates of infection with *Pseudomonas aeruginosa*, *Staphylococcus aureus*, *Stenotrophomonas sp.*, *Haemophilus influenzae*, and *Aspergillus sp.*.

Cluster Analysis

Between 3-5 clusters were defined for each combination of variables using Partitioning Around Medoids (PAM) cluster analysis. The iterative cluster methods meant that 1981 cluster models were developed, and while it would be advantageous to calculate the optimal cluster number for every model using a cluster index (such as silhouette width or elbow method), and then cluster every model based on its optimal cluster number, these methods would be drastically limited by computer processing time. Therefore, instead of choosing the optimal number of clusters for a single data set, the dataset that was optimal for the small range of clusters was identified.

In detail, missing values were excluded from each combination of variables, variables were normalised between 0-1, and Euclidean distance was calculated as the measure of dissimilarity between all clinical encounters. The average silhouette width, a measure of within cluster similarity and between cluster dissimilarity, of each cluster model was ranked. The models with the highest silhouette widths were selected for visualisation using t-SNE plots (a dimensionality reduction technique) [1].

In total, 5943 cluster models were developed (1981 models per cluster number), which were composed of between 12467 – 31218 encounters comprising 525 – 681 individuals. The models

ranged widely in silhouette widths and t-SNE plots, in which variable number was found to strongly influence the quality of clusters. Higher numbers of variables included in the models resulted in robust t-SNE plots with lower silhouette widths compared to models with low variable numbers (Figure S1).

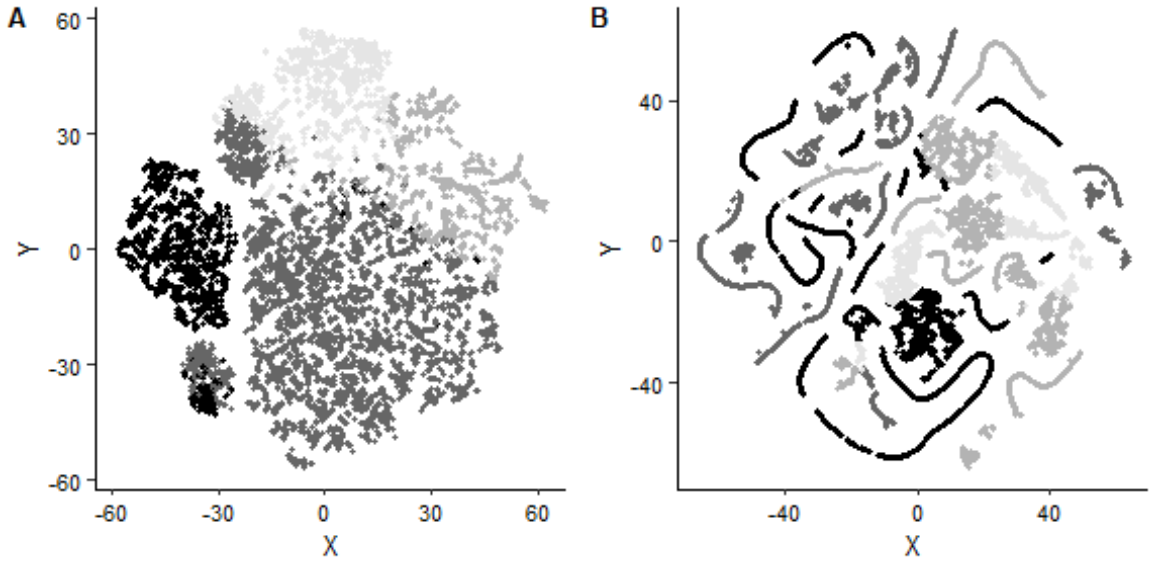

Figure S1. t-SNE plots of A) the optimal model with good cluster distinctions (9 variables: body mass index (BMI), height, hospitalisations in prior year, cough, and previous rates of infection with *Pseudomonas aeruginosa*, *Staphylococcus aureus*, *Stenotrophomonas sp.*, *Haemophilus influenzae*, and *Aspergillus sp*) and B) a poor cluster model with disjointed clusters (4 variables: BMI, cough, OPEx, *Aspergillus sp.*)

Time to Event Analyses

Time-to event analyses were conducted using a Cox proportional hazards regression [2]. Specifically, a marginal means and rates model was used for risk of recurrent PEx and hospitalization events [3], and a standardized survival model was used to calculate risk of death and transplant from an individual's first cluster assignment [4]. The analyses were carried out on the top 36 models identified from silhouette widths and t-SNE plots. The outcome analysis also

varied widely across models, where higher numbers of variables contributed to better models (lower Bayesian information criterion (BIC)) on average. Strong separation in mild outcomes (time-to hospitalisation and time-to PEx treated with oral antibiotics) were prioritised over a strong separation in severe outcomes (time-to death and time-to transplant).

Optimal Cluster Model

Table S1. Description of clinical variables and patient characteristics of encounters included in the optimal cluster model; mean(SD) unless otherwise stated.

Variable	Mean (SD)	Range
Age	10.79 (4.38)	2 - 18
BMI Z Score	-0.31 (1.04)	-9.15 - 4.38
Height Z Score	-0.44 (1.02)	-5.41 - 3
Weight Z Score	-0.49 (1.11)	-7.58 - 4.15
P. aeruginosa	0.18 (0.27)	0 - 1
S. aureus	0.33 (0.23)	0 - 1
B. cepacia complex	0.01 (0.08)	0 - 1
Achromobacter sp.	0.01 (0.05)	0 - 0.94
Aspergillus sp.	0.09 (0.16)	0 - 1
H. influenzae	0.09 (0.11)	0 - 1
Stenotrophomonas sp.	0.04 (0.1)	0 - 1
Methicillin Resistant *S. aureus*	0.01 (0.06)	0 - 1
PEx treated with IV antibiotics in Prior Year	0.37 (0.82)	0 - 11
PEx Treated with Oral Antibiotics in Prior Year	1.19 (1.28)	0 - 8
Hospitalisations in Prior Year	0.46 (0.96)	0 - 10
Ontario Marginalisation Index	2.34 (1.22)	1 - 5
Age at Diagnosis	1.44 (2.46)	0 - 16.3
Cough	3.02 (1.16)	1 - 5
FEV_1 % Predicted	79.29 (21.13)	16.26 - 146.58
Class I-III n(%)	11248 (92.2)	
Female n(%)	6370 (52.2)	
PI n(%)	11205 (91.8)	
White n(%)	10845 (88.9)	
Ivacaftor n(%)	194 (1.6)	
CFRD n(%)	546 (4.5)	
Chronic Inhaled Antibiotics n(%)	2591 (21.2)	

Cluster prediction of Future FEV_1

Table S2. Coefficients and confidence intervals for the predicted rate of change in FEV_1 % predicted over 1 year stratified across clusters.

	Time	Age	Time * Age	SD
Cluster A	-3.36 (-7.57 - 0.84)	-1.72 (-2.08 - -1.36)	0.15 (-0.17 - 0.047)	0.97
Cluster B	0.81 (-1.47 - 3.10)	-1.61 (-1.81 - -1.42)	-0.08 (-0.25 - 0.10)	7.08
Cluster C	-5.77 (-9.47 - -2.06)	-0.17 (-0.58 - 0.25)	0.26 (0.00 - 0.51)	6.09
Cluster D	13.24 (8.75 - 17.73)	-2.14 (-2.55 - -1.74)	-0.67 (-0.99 - -0.35)	9.64

Internal Validation

Using a K-Nearest Neighbours approach, Euclidean distance between new data and the centre of each cluster is determined to identify which cluster the new data resembles most. This was carried out using a Nearest Neighbours kd-tree searching algorithm [5].

GOSH Data Exclusions

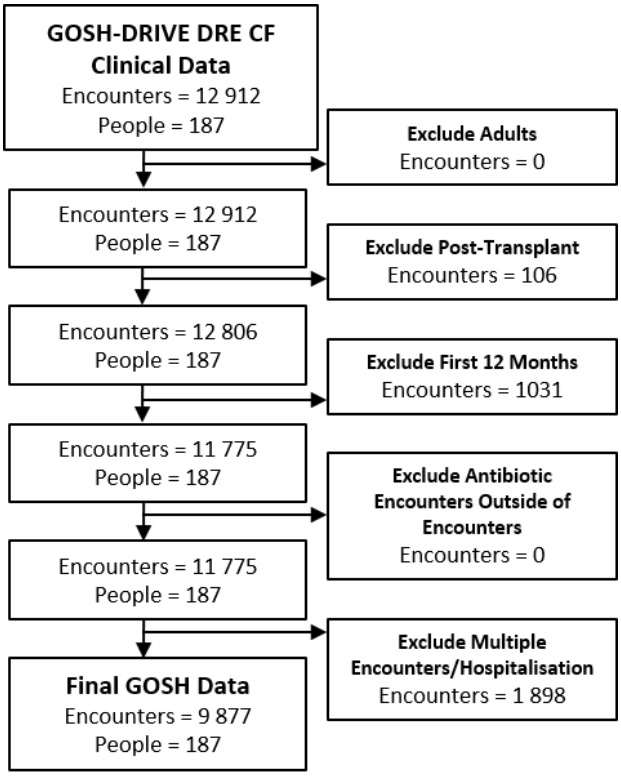

Figure S2. Description of data exclusions for the GOSH Data

External Validation

Table S3. Hazards ratios and confidence intervals for each cluster as compared to cluster A across the GOSH and Revised TCF validation time-to hospitalisation analysis. Bold values are significant ($p < 0.05$)

	Hospitalisation GOSH	Hospitalisation Revised TCF
Cluster B	**2.15 (1.15-4.02)**	1.28 (0.81-2.03)
Cluster C	**3.64 (2.01-6.59)**	1.51 (0.91-2.51)
Cluster D	**6.13 (4.16-9.02)**	**3.97 (2.45-6.42)**

References

1. Donaldson J. T-Distributed Stochastic Neighbor Embedding for R (t-SNE). 2016.

2. Therneau TM, Lumley T. Survival Analysis. 2019.

3. Amorim LD, Cai J. Modelling recurrent events: A tutorial for analysis in epidemiology. *International Journal of Epidemiology* 2015; 44: 324–333.

4. Sykes J, Stanojevic S, Goss CH, Quon BS, Marshall BC, Petren K, Ostrenga J, Fink A, Elbert A, Stephenson AL. A standardized approach to estimating survival statistics for population based cystic fibrosis registry cohorts. *J Clin Epidemiol* 2016; 70: 206–213.

5. Beygelzimer A, Kakadet S, Lanhford J, Arya S, Mount D, Li S. Fast Nearest Neighbor Search Algorithms and Applications. 2019.