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

Nicole Filipow[1,2], Gwyneth Davies[1,5], Eleanor Main[1], Neil J Sebire[1,5], Colin Wallis[5], Felix Ratjen[2,3,4], Sanja Stanojevic[2,6]

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

Word Count

Abstract [241]
Manuscript [3584]

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.

Abstract

Background Cystic Fibrosis (CF) is a multisystem disease in which assessing disease severity based on lung function alone may not be appropriate. The aim of the study was to develop a comprehensive machine-learning algorithm to assess clinical status independent of lung function in children.

Methods A comprehensive prospectively collected clinical database (Toronto, Canada) was used to apply unsupervised cluster analysis. The defined clusters were then compared by current and future lung function, risk of future hospitalisation, and risk of future pulmonary exacerbation (PEx) treated with oral antibiotics. A K-Nearest Neighbours (KNN) algorithm was used to prospectively assign clusters. The methods were validated in a paediatric clinical CF dataset from Great Ormond Street Hospital (GOSH).

Results The optimal cluster model identified four (A-D) phenotypic clusters based on 12,200 encounters from 530 individuals. Two clusters (A,B) consistent with mild disease were identified with high FEV_1, and low risk of both hospitalisation and PEx treated with oral antibiotics. Two clusters (C,D) consistent with severe disease were also identified with low FEV_1. Cluster D had the shortest time to both hospitalisation and PEx treated with oral antibiotics. The outcomes were consistent in 3,124 encounters from 171 children at GOSH. The KNN cluster allocation error rate was low, at 2.5% (Toronto), and 3.5% (GOSH).

Conclusion Machine learning derived phenotypic clusters can predict disease severity independent of lung function and could be used in conjunction with functional measures to predict future disease trajectories in CF patients.

Introduction

Cystic fibrosis (CF) is characterized by lung disease, pancreatic insufficiency (PI), malabsorption of nutrients, and can lead to numerous comorbidities such as CF-related diabetes (CFRD) and male infertility[1, 2]. Respiratory complications are the greatest cause of mortality, and therefore, current standards for assessing disease severity, monitoring disease progression, and evaluating clinical trials rely heavily on lung function as an outcome measure[3].

People with CF have seen profound improvements in care, and many children now maintain lung function in the normal range[4–6]. Novel therapeutics (e.g. ivacaftor, and elexacaftor/tezacaftor/ivacaftor) that correct the underlying molecular defect responsible for CF are expected to further improve lung function decline and the overall prognosis of people living with this disease[7]. Nonetheless, these treatments are not a cure, and there is still a need to monitor disease progression, and thus a call to develop new measures that adequately detect mild disease and can predict disease trajectories [8], especially in the paediatric age group.

Unsupervised cluster analysis, a form of machine learning, is a common approach to identify subgroups of disease. Unlike supervised methods, where multivariate classification is anchored to pre-defined labels, such as death or arbitrary thresholds of lung function, unsupervised analysis will group data based on natural patterns found both within and between variables[9]. Relevance of the groups, or clusters, are then assessed though association with outcome measures. The method has been applied to respiratory illnesses including COPD, asthma, and bronchiectasis[10–12], as well as in CF[13–15], however these studies have associated clusters with death and transplant, which are uncommon events in paediatrics.

In order to develop a paediatric-specific outcome measure, it is important that a stronger link is established between routinely collected clinical variables and milder outcomes. Furthermore, any new measure should be derived independently of lung function in order to deviate from its historic reliance, and accordingly provide a complementary measure to monitor CF disease.

The aims of this study were to: 1) Use unsupervised clustering to identify clusters in a large paediatric Toronto CF dataset (TCF), and investigate whether these clusters distinguish patients based on current and future lung function measured by spirometry (forced expiratory volume in 1 second (FEV_1)), risk of future hospitalisation, and risk of future pulmonary exacerbation (PEx) treated with oral antibiotics. 2) Validate the clusters internally by investigating trends across age and time. 3) Validate the clusters externally using a large paediatric UK CF dataset from Great Ormond Street Hospital (GOSH). 4) Evaluate the repeatability in applying the clusters as a clinical measure.

Methods

Data

The TCF is an encounter-based registry that records clinical data at every CF clinic visit. To ensure relevance to the current CF population, analyses were limited to the most recent two decades (2000-2018). Adults (> 18 years) were excluded, and clinical data recorded after a lung transplant were censored. Oral and inhaled antibiotics captured outside of clinical encounters were recorded but not included as an encounter, and each hospitalisation was summarised as a single encounter. The recent history of hospitalisations and PEx events treated with antibiotics were captured using a 12-month look-back window. Missing data were random throughout the dataset and were excluded from the cluster model.

Clinical data from patients with CF from a second specialist children's hospital, GOSH, were used to validate the TCF derived clusters. Data were obtained from hospital admission records, microbiology lab results, spirometry tests and clinical notes, which were available from 2009-2017. Data were merged and analysed in the GOSH-DRIVE digital research environment (DRE) (an electronic healthcare records (EHR) database) (DRE, Aridhia Inc, Edinburgh, UK). The same exclusions applied to the TCF dataset were also applied to the GOSH data. This study was approved by the Research Ethics Board at the Hospital for Sick Children (REB#1000060824) and covered under the ethical approval 17/LO/0008 (R&D#19IA07) at GOSH.

Cluster Analysis

All analyses were carried out in R software[16]. Paediatric CF physician input and the CF literature identified an initial list of 25 variables as relevant to CF health (**Table 1**). In order to reduce noise and redundancy in the model, Pearson correlation tests and principal component analyses were used to inform decisions on excluding correlated variables and those with minimal contribution to the variance in the data. FEV_1 was deliberately excluded from the model and was instead assessed as an outcome measure to corroborate the disease severities of the clusters.

Partitioning Around Medoids (PAM) clustering[17] was used to generate between 3-5 clusters. Initially, clustering was carried out on all combinations of variables (range: 3-11), resulting in a total of 1981 cluster models per cluster number. The maximum number of variables included in the cluster combinations were restricted to 11 in the first instance for computational and practical reasons. Superior models were identified by silhouette width: a measure of within-cluster similarity. Additional details are provided in the Online Supplement.

Outcomes

The final candidate models (n=36) were assessed by comparing between-cluster differences in outcomes. Specifically, the models were ranked by model fit estimated from the Bayesian information criterion (BIC) of each of time-to hospitalisation (typically courses of IV antibiotics), time-to PEx treated with oral antibiotics, and a linear regression model of FEV_1% predicted (calculated from GLI reference equations[18]) (see Online Supplement for details). The models were also ranked by sample size. Those that ranked best across all four parameters were independently assessed, and an optimal model was chosen as the one with the best between-cluster separation in outcomes.

The final optimal model was also analysed to determine cluster association with future lung function as the rate of change in FEV_1% predicted at one year from each encounter (estimated using linear mixed models with random slopes and intercepts) stratified between clusters [19]. The proportion of encounters in each cluster was also calculated for different thresholds of FEV_1% predicted. Finally, time-to transplant or death from an individual's first cluster assignment.

Internal Validation

On average, older children in the study cohort were anticipated to be more unwell than younger children, and children of the same age were anticipated to be healthier in the late 2010's than in the early 2000's. To validate the cluster detection of these trends in disease severity, the proportion of encounters in each cluster were assessed across time and age.

External Validation

Clusters were defined for the GOSH data using the variables identified in the TCF optimal model. Clustering was also carried out on a smaller TCF dataset using the same time period as the GOSH data for a matched comparison. Between-cluster trends in outcomes were compared between the populations.

Cluster Allocation

To apply the clusters as a clinical measure, new data (i.e. from a clinic encounter) can be allocated to the closest cluster using a KNN algorithm[20]. The mean error rate of the method was estimated by assigning a randomly selected 20% of encounters (test data) to clusters generated from the remaining 80% of data[21], which was compared to the cluster assignment from the optimal model for both TCF and GOSH datasets, in 1000 iterations.

Results

Data

The TCF contains 78,014 clinical encounters from 1,309 people with CF. After exclusions, the dataset included 20,586 encounters from 575 children between 2000-2018 (**Figure 1**). From the initial list of 25 variables (**Table 1**), the review process identified 11 candidate variables for iterative clustering (See Online Supplement for variable selection details).

Table 1. Description of the initial list of TCF variables identified for their relevance to CF Health

Group	Variable	Missing Encounters (%)	Class	Definition
Anthropometry	BMI	22.2	Numeric	Z Scores calculated from the CDC Growth Charts [22]
	Height	10.6		
	Weight	18.7		
Microbiology	*Pseudomonas aeruginosa*	11.7	Numeric	Proportion of positive cultures detected from all previous cultures ever taken
	Staphylococcus aureus	11.7		
	Burkholderia cepacia complex	11.7		
	Achromobacter sp.	11.7		
	Aspergillus sp.	11.7		
	Haemophilus influenzae	11.7		
	Stenotrophomonas sp.	11.7		
	Methicillin Resistant *S. aureus* (MRSA)	11.7		
Hospitalisations & Pulmonary Exacerbations (PEx)	PEx treated with IV antibiotics in prior year	0.0	Numeric	Number of events in previous 12 month rolling window
	PEx treated with oral antibiotics in prior year	0.0		
	Hospitalisations in prior year	0.0		
Demographics	Sex	0.0	Categorical / Binary	Male / Female
	Ontario Marginalisation Index (as a proxy for socioeconomic status) [23]	6.4	Numeric (Integer)	1 - 5, where a high number means more deprived
	Age	0.0	Numeric	Years
	Age at Diagnosis	0.0		
	Ethnicity	0.3	Categorical	White / Black / NE Asian / SE Asian / Other
Genetics	Functional Class	1.7	Categorical	I-III / IV-V
Symptoms / Comorbidities	Cough	32.1	Numeric (Integer)	Reported by clinician based on patient symptoms at encounter 1 = None, 2 = Only with Therapy, 3 = Occasional, 4 = Increased, 5 = Chronic
	Pancreatic Insufficiency	0.1	Categorical / Binary	Pancreatic Insufficient / Pancreatic Sufficient; Determined from prescription of pancreatic enzymes
	CFRD	0.0	Categorical / Binary	Yes / No
Treatments	Ivacaftor	0.0	Categorical / Binary	Yes / No
	Inhaled Antibiotics (as a proxy for *P. aeruginosa* infection)	0.0	Categorical	Chronic / Intermittent / Never

Optimal Cluster Model

The final optimal cluster model consisted of 4 clusters comprised of 9 variables: body mass index (BMI), height, hospitalisations in prior year, cough, as well as previous rates of infection with *Pseudomonas aeruginosa*, *Staphylococcus aureus*, *Stenotrophomonas sp.*, *Haemophilus influenzae*, and *Aspergillus sp*. The model was generated from 12,200 complete encounters involving 530 individuals aged 2-18 years (mean=10.79, SD=4.38). Of the individuals included, 11% received a transplant and 12% died before the end of the study period. Thirteen individuals received ivacaftor, making up 0.03% of the encounters (See Online Supplement **Table S1** for additional patient characteristics).

Based on the assessment of individuals within the clusters, two clusters were consistent with more mild disease (Clusters A&B) and two clusters were consistent with more severe disease (Clusters C&D). The four-cluster model provided more granularity on between-cluster outcomes than the three-cluster model, while the five-cluster model did not show a meaningful distinction.

The addition of FEV_1 to the optimal model did not change the results, suggesting the variables included already combine to influence FEV_1.

Cluster Characteristics

Both severe clusters (C&D) were characterized by low BMI, reduced height and weight, high numbers of hospitalisations in prior year, high rates of previous *Aspergillus sp.* infection, and high prevalence of chronic cough. These clusters were composed of older children with a large prevalence of both PI and CFRD, and a high use of chronic inhaled antibiotics (**Table 2**). In addition, cluster C had the highest rate of previous *P. aeruginosa* infection, and cluster D had the greatest number of PEx treated with oral antibiotics in prior year.

In contrast, the two mild clusters (A&B) were composed of younger children with lower prevalence of CFRD and PI, and were characterized by high growth parameters, low rates of previous *P. aeruginosa* and *Aspergillus sp.* infection, and low numbers of hospitalisations in prior year (**Table 2**). Cluster A also had the lowest prevalence of chronic cough.

Table 2. Summary of patient characteristics, and clinical variables for each cluster

	Variable	Cluster A	Cluster B	Cluster C	Cluster D
Count	Encounters (n)	2028	6583	1732	1857
	People (n)	356	414	147	307
Nine Clustering Variables	BMI Z Score[a]	-0.03 (-3.59 - 4.26)	-0.21 (-9.15 - 4.38)	-0.69 (-7.15 - 2.11)	-0.65 (-8.42 - 2.47)
	Height Z Score[a]	-0.23 (-3.67 - 2.93)	-0.35 (-3.61 - 3.00)	-0.70 (-3.46 - 2.27)	-0.72 (-5.41 - 2.59)
	P. aeruginosa[a]	0.07 (0.00 - 0.64)	0.08 (0.00 - 0.50)	0.77 (0.22 - 1.00)	0.11 (0.00 - 0.69)
	S. aureus[a]	0.31 (0.00 - 1.00)	0.34 (0.00 - 1.00)	0.26 (0.00 - 1.00)	0.42 (0.00 - 1.00)
	Stenotrophomonas sp.[a]	0.02 (0.00 - 0.86)	0.04 (0.00 - 1.00)	0.04 (0.00 - 0.85)	0.09 (0.00 - 1.00)
	H. influenzae[a]	0.09 (0.00 - 0.67)	0.11 (0.00 - 1.00)	0.03 (0.00 - 0.55)	0.10 (0.00 - 0.82)
	Aspergillus sp.[a]	0.03 (0.00 - 0.81)	0.07 (0.00 - 1.00)	0.18 (0.00 - 0.90)	0.15 (0.00 - 1.00)
	Hospitalisations in Prior Year[a]	0.16 (0.00 - 6.00)	0.28 (0.00 - 8.00)	0.77 (0.00 - 9.00)	1.11 (0.00 - 10.00)
	Cough[a]	1.10 (1.00 - 2.00)	3.03 (2.00 - 4.00)	3.38 (1.00 - 5.00)	4.75 (3.00 - 5.00)
Variables Excluded from Cluster Model	Age[a]	8.66 (2.00 - 17.99)	10.23 (2.00 - 18.00)	13.60 (2.01 - 18.00)	12.48 (2.04 - 18.00)
	B. cepacia complex[a]	0.00 (0.00 - 0.80)	0.02 (0.00 - 1.00)	0.00 (0.00 - 0.43)	0.02 (0.00 - 1.00)
	Achromobacter sp.[a]	0.00 (0.00 - 0.93)	0.01 (0.00 - 0.63)	0.00 (0.00 - 0.92)	0.03 (0.00 - 0.94)
	MRSA[a]	0.01 (0.00 - 1.00)	0.01 (0.00 - 1.00)	0.00 (0.00 - 0.43)	0.02 (0.00 - 1.00)
	Ontario Marginalisation Index[a]	2.34 (1.00 - 5.00)	2.31 (1.00 - 5.00)	2.31 (1.00 - 5.00)	2.44 (1.00 - 5.00)
	PEx Treated with Oral Antibiotics in Prior Year[a]	0.84 (0.00 - 8.00)	1.20 (0.00 - 8.00)	0.94 (0.00 - 7.00)	1.78 (0.00 - 7.00)
	PEx Treated with IV Antibiotics in Prior Year[a]	0.11 (0.00 - 4.00)	0.24 (0.00 - 5.00)	0.62 (0.00 - 11.00)	0.91 (0.00 - 8.00)
	Weight Z Score[a]	-0.17 (-5.18 - 4.13)	-0.36 (-6.12 - 4.15)	-0.90 (-6.20 - 2.24)	-0.90 (-7.58 - 2.98)
	Age at Diagnosis[a]	1.34 (0.00 - 15.00)	1.38 (0.00 - 15.00)	1.56 (0.00 - 15.89)	1.64 (0.00 - 16.30)
	Class I-III (%)	90.34	92.25	93.19	93.11
	Female (%)	50.20	49.64	56.47	59.56
	PI (%)	86.79	91.28	98.27	93.38
	CFRD (%)	1.92	2.61	10.68	8.08
	White (%)	88.76	89.73	87.93	86.97
	Ivacaftor (%)	3.16	1.47	0.58	1.24
	Chronic Inhaled Antibiotics (%)	10.60	16.63	53.70	18.90

[a]Mean (Range)
Variable definitions are found in Table 1

Cluster Outcomes

There were 10351 complete lung function measurements in the optimal model. Cluster A had the highest FEV_1 (mean(SD) = 93% (15.5%)), Cluster B had intermediate FEV_1 (mean(SD) = 83.5% (17.9%)), and Cluster C & Cluster D had low FEV_1 (mean(SD) = 68.5% (21%); 64.6% (22.2%)) (**Figure 2A**). Differences in FEV_1 were significant between all clusters (post-hoc Tukey: $p < 0.05$). Lung function alone did not distinguish cluster (**Figure 2A**).

The greatest percent of encounters in cluster A (29.6%) had a FEV_1 >100% and the greatest percent of encounters in cluster D (46.6%) had a FEV_1 <40% (**Figure 2B**). For encounters in cluster A with FEV_1 <40% ($n_{encounter} = 6$, $n_{people} = 4$), inconsistent data entry was likely responsible when the trajectory of each individual's encounters were investigated.

FEV_1 in cluster A declined less with increasing age, and in cluster B FEV_1 was relatively stable (**Figure 2C**). Cluster C had the steepest decline in FEV_1 over 1 year, which remained stable across age (**Figure 2C**). FEV_1 in cluster D increased over 1 year, which was more profound in early childhood. The slope in this cluster was highly variable between individuals (SD = 9.64) (**Figure 2C**).

The risk of both future hospitalisation and future PEx treated with oral antibiotics was lowest in cluster A, and highest in cluster D (**Figure 3A,B; Table 3**). Cluster B had a slightly higher risk of PEx treated with oral antibiotics than cluster C, while cluster C had a comparably much higher risk of hospitalisation (**Figure 3A,B; Table 3**). While death and transplant in childhood (<18) were rare (n=21), time to death or transplant later in adulthood was linked to first severe disease cluster in childhood, in which Cluster D had the highest risk of both death and transplant (**Figure 3A,B; Table 3**).

Table 3. Hazards ratios and confidence intervals for each cluster as compared to cluster A across each time-to event analysis. Bold values are significant ($p < 0.05$)

	PEx Treated with Oral Antibiotics	Hospitalisation	Death	Transplant
Cluster B	**1.87 (1.66-2.11)**	**2.09 (1.63-2.7)**	4.55 (0.62-33.66)	1.04 (0.3-3.56)
Cluster C	**1.81 (1.46-2.25)**	**4.15 (2.9-5.96)**	5.05 (0.68-37.75)	2.04 (0.61-6.79)
Cluster D	**2.31 (1.94-2.74)**	**7.44 (5.58-9.91)**	**14.81 (1.91-114.78)**	**4.8 (1.35-17.03)**

Internal Validation

Age and year of encounter were not included in the model, nonetheless, there were clear age-related and temporal trends, such that older children were more likely to be in the more severe clusters C&D, and newer cohorts (at the same age) were more likely to be in the milder clusters A&B (**Figure 4**). These observations were consistent with expected temporal and age-related trends.

External Validation

The GOSH data included 12,912 encounters from 187 children. Cough and oral antibiotic data were not available in this dataset so were removed from analysis. After data exclusions and the removal of missing values, the cluster model comprised 3,124 encounters from 171 children aged 1-17.9 (mean=8.2) (See Supplement **Figure S2** for exclusion criteria).

For direct comparison, the TCF data were reanalysed using the same time period and variables as the GOSH data; there were 6,623 encounters from 338 children aged 2-18 (mean=11.1) in the revised TCF cluster model.

There was a similar gradient in risk of hospitalisation and FEV_1 in both populations. Cluster D had the shortest time to hospitalisation, and the lowest FEV_1% predicted (mean (SD): GOSH:

69.2% (19.3%), TCF: 68.5% (21.7%)), whereas cluster A had the lowest risk of hospitalisation and highest FEV_1% predicted (mean (SD): GOSH: 85.7% (16.4%), TCF: 87.8% (18%)) (**Figure 5**). Cluster differences in FEV_1 were all significant (post-hoc Tukey: $p < 0.05$), with the exception of clusters C and D in the GOSH analysis ($p = 1.00$).

Cluster Allocation

The KNN method for cluster allocation accurately assigned the test encounters to the same clusters as the original clustering, with an error rate of 2.5% for TCF encounters and 3.5% for GOSH encounters.

Discussion

Phenotypic clusters using a range of clinical outcomes collected during routine clinical care allow for a comprehensive overview of CF health, and our results show they meaningfully represent both mild and severe classes of CF disease. Within clusters, lung function was concordant with disease outcomes, whereas the range of individual lung function values observed within each cluster was wide. In the current era of CF, where up to 85% of affected children are reported to have mild to normal lung function [4–6], a multifactorial tool may provide further insight into disease progression.

The unsupervised algorithm benefits from the exclusion of FEV_1 since young children who cannot perform spirometry are still captured. Where FEV_1 is normal, the clusters may aid to explain who is at a greater risk of hospitalisation or PEx and who could benefit from targeted management.

It was surprising that FEV_1 in cluster D increased over 1 year in younger children, although it suggests that clinicians may already recognise that these individuals have more severe disease and require more intense interventions. The greater variability in the slopes over 12 months further suggests that these trends may reflect treatment effects.

The validation carried out in the GOSH population was very similar to the results of the TCF data re-analysed with a subset of the original population (n_{people} = 338). Although the GOSH population was at a higher risk of hospitalisations overall, a further investigation into clinical practices reveals differences between these two centres. For instance, routine regular hospitalisations for IV antibiotics are common for children with severe disease at GOSH, whereas hospitalisations in Toronto are typically only for acute exacerbations. Despite this significant difference in management approach, FEV_1 values were similar across clusters between the two populations. Clustering on the same variables yielded similar outcomes, highlighting the robustness of the cluster method. To ensure applicability in the modern CF era, the model should be updated in a predominantly new-born screening cohort, as well as in genetically diverse cohorts to ensure generalizability.

Multivariate scores have rarely been used in routine clinical care, but as EHR become more common and data are stored centrally, the implementation of this type of phenotypic clustering into clinical practice will become more feasible. The KNN algorithm can be used to calculate a cluster for each CF encounter based on the input of the nine cluster variables (see Table 2). As such, following appropriate governance and evaluation, the algorithm could be incorporated into EHR systems to provide clinicians with an overall picture of patient status and inform clinical decision making. Investigation into treatment effects across clusters with chronic therapies (e.g. dornase alfa and hypertonic saline) were limited in this study by missing and inconsistent

data but need to be explored further to better understand whether treatments can modify disease severity status.

The algorithm could also be implemented in patient portals to their EHR or apps to provide patients with a more comprehensive picture of their health status where single clinical measures, such as BMI or recent infection, may not be as meaningful or interpretable indicators of health. Future work involving patients/families will highlight the appropriate presentation of cluster labels to better provide insight into overall health status.

Additional potential application use cases of clusters are as an endpoint in research studies where changes in FEV_1 are not detectable; improvements in health status may instead be indicated by movement from a severe cluster (C/D) to a mild cluster (A/B). This may be particularly attractive in trials involving young children, and/or in the rapidly evolving era of highly effective CFTR modulator treatments being integrated into routine care. Clusters may also be used in national registry reports to standardise clinics or regions by disease severity for matched comparisons of populations, or to highlight those who may benefit from more resources.

A current barrier to implementation is the issue of missing clinical data, in which future analyses are necessary to triangulate evidence from existing data to ensure robust estimation of clusters. Future work will also explore the integration of filters to prevent the mis-calculation of clusters from unrealistic clinical values. A limitation of the approach is that we standardised variables, which means the relative importance of each variable in defining the transition between cluster is unknown, and future work is needed to explore this.

Cluster analysis involves largely subjective decisions each step of the cluster pipeline, and further refinement of the model may include the selection of different parameters to better

understand the stability of the model. At present, the optimal model has a good performance and is an important first step, but it may not be the best model as highly effective CFTR modulators begin to change clinical outcomes and prognosis. Future works needs to determine how to routinely update models as patient characteristics and available data changes. For instance, cough was repeatedly highlighted as important in defining outcomes during iterative clustering but did not appear to impact the GOSH validation. While cough is limited by subjectivity and a high proportion of missing values, its inclusion allowed the model to encompass patient-reported symptoms. More objective symptoms collected routinely through EHR or patient apps may be required to better capture the lived experience of the patient. CFTR modulator therapy was excluded from the model due to minimal data but should be considered once these treatments are available to a wider proportion of CF patients. Similarly, lung clearance index (LCI) should be considered for inclusion as it becomes a more routinely collected clinical variable, and extra-pulmonary manifestations not captured in the registry should be included when using EHR data. An advantage of using an integrated cluster algorithm means that the variables can be updated, the time period adjusted, and sub-groups refined as new information becomes available.

This analysis focused on time-dependent, continuous variables to assess cluster change over time, but also because clustering with partitioning around medoids (PAM) is more favourable using continuous variables. For example, the inclusion of genetic sex resulted in clusters completely defined by sex alone, with less association with clinical outcomes. This demonstrates that the unsupervised nature of clustering, in which the algorithm aims to find similar groups without a purpose, requires clinical interpretation to ensure real-world value. In selecting meaningful, continuous variables, the look-back window to capture an individual's clinical

history means the frequency of microbiology sampling could introduce a potential bias, and sicker patients with more hospitalisations may be overrepresented.

The phenotypic clusters were broadly comparable to a previous study of an adult CF population, which found *P. aeruginosa*, IV antibiotics, and pancreatic insufficiency were consistent in severe clusters with a high risk of death, and vice versa for milder clusters with a low risk of death [13]. The study identified 7 clusters within 25 variables. We restricted the number of possible clusters from 3-5 to simplify the practical interpretation of the results, and we were able to reduce the number of variables to 9 for a more feasible clinical application. Despite these differences, both approaches identified robust clusters with similar variables associated with severe disease. Other cluster analyses in CF have anchored clusters to physician determined levels of disease severity, which may be biased by indication [24], or have included FEV_1 in the model [14]. Our analysis provides a more objective categorization of disease severity, by statistically testing the clusters against known indicators of health that are mainly excluded from the model itself.

Cluster analyses have been applied to other diseases and further emphasize that knowledge-based inclusion of variables is necessary to ensure meaningful translation of the results [11, 12, 25, 26]. Another common approach has been to transform variables into linear combinations [10, 11, 25], which we decided would complicate the interpretation of our results. Two studies have also gone further to prospectively apply the clusters by generating an external scoring system using decision tree analyses [13, 25]. The advantage of the KNN algorithm is that it directly allocates clusters to an encounter to predict disease severity. Our results show it has an extremely low error rate (< 5%).

Multivariate clinical scoring systems have been derived in CF, but they typically assume an additive/multiplicative independent association between the variables included [27–30]. Cluster

analysis instead makes no assumption about the nature of the relationship, only how individuals are similar to one another through shared patterns across the variables. Additionally, multivariate models must be linked to an event of interest or a specific outcome – and the advantage of this approach is that the clusters are correlated with important clinical outcomes, but not developed with an explicit prediction model of a single outcome.

Conclusion

It is feasible to develop machine learning based phenotypic clusters that summarize the overall health status of children living with CF, and these may provide a holistic way to track the progression of disease across childhood. The cluster algorithm can be updated regularly to accompany a rapidly changing therapeutic environment.

Figure Legends

Figure 1. Flow chart summarizing the study population and data exclusions applied to the TCF dataset.

Figure 2. The association between each cluster and lung function: **A)** Violin plots summarizing FEV_1% predicted by cluster, where the violin height represents the density of data. Boxes within the violins display the median (middle line) and 25% and 75% quartiles, horizontal lines indicate 1.5 x interquartile range, and points are outliers. **B)** Stacked Bar graph demonstrating the proportion of encounters in each cluster across different thresholds of FEV_1% predicted. **C)** The predicted rate of change in FEV_1 % predicted over 1 year stratified across clusters. The predicted trajectory for three different ages (10,13,16) are shown. The corresponding slopes are provided in Table S2 of the Online Supplement.

Figure 3. Time to event analyses (marginal means and rates model) by cluster. Analysis included repeated events where time was re-initiated when an individual switched clusters for **A)** Time to PEx treated with oral antibiotics, and **B)** Time to hospitalisation. Analysis also included time to event from an individual's first cluster assignment in the TCF for **C)** Time to death, and **D)** Time to transplant. Hazard ratios are provided in Table 3.

Figure 4. A) The proportion of individuals in each cluster across age at different years and **B)** The proportion of individuals in each cluster across time at different ages. There were 7494 encounters before 2010, and 4706 encounters after 2010. The proportion of encounters in Cluster A and Cluster D increased across decades, and the proportion of encounters in Cluster B and Cluster C decreased across decades.

Figure 5. Comparison of cluster outcomes in the GOSH **(i)** and Toronto **(ii)** datasets. **A)** Time to hospitalisation (marginal means and rates model) by cluster. Analysis included repeated events where time was re-initiated when an individual switched clusters. Hazard ratios are provided in Table S3 of the Online Supplement. **B)** Violin plots summarizing FEV_1% predicted by cluster, where the violin height represents the density of data. Boxes within the violins display the median (middle line) and 25% and 75% quartiles, horizontal lines indicate 1.5 x interquartile range, and points are outliers.

References

1. Ronan NJ, Elborn JS, Plant BJ. Current and emerging comorbidities in cystic fibrosis. *Presse Medicale* 2017; 46: e125–e138.

2. Jackson AD, Goss CH. Epidemiology of CF: How registries can be used to advance our understanding of the CF population. *Journal of Cystic Fibrosis* 2018; 17: 297–305.

3. Szczesniak R, Heltshe SL, Stanojevic S, Mayer-Hamblett N. Use of fev1 in cystic fibrosis epidemiologic studies and clinical trials: A statistical perspective for the clinical researcher. *Journal of Cystic Fibrosis* 2017; 16: 318–326.

4. Cystic Fibrosis Trust. UK Cystic Fibrosis Registry Annual Data Report 2018. 2019 p. 1–87.

5. Cystic Fibrosis Foundation. 2018 Patient Registry Annual Data Report. 2019 p. 1–74.

6. Cystic Fibrosis Canada. The Canadian Cystic Fibrosis Registry 2018 Annual Data Report. 2019 p. 1–43.

7. Bell SC, Mall MA, Gutierrez H, Macek M, Madge S, Davies JC, Burgel PR, Tullis E, Castaños C, Castellani C, Byrnes CA, Cathcart F, Chotirmall SH, Cosgriff R, Eichler I, Fajac I, Goss CH, Drevinek P, Farrell PM, Gravelle AM, Havermans T, Mayer-Hamblett N, Kashirskaya N, Kerem E, Mathew JL, McKone EF, Naehrlich L, Nasr SZ, Oates GR, O'Neill C, et al. The future of cystic fibrosis care: a global perspective. *The Lancet Respiratory Medicine* 2020; 8: 65–124.

8. De Boeck K, Lee T, Amaral M, Drevinek P, Elborn JS, Fajac I, Kerem E, Davies JC. Cystic fibrosis drug trial design in the era of CFTR modulators associated with substantial clinical benefit: stakeholders' consensus view. *Journal of Cystic Fibrosis* Elsevier B.V. 2020; 19: 688–695.

9. Everitt BS, Landau S, Leese M, Stahl D. Cluster Analysis. 5th ed. London: John Wiley & Sons, Ltd; 2011.

10. Burgel P-R, Paillasseur J-L, Caillaud D, Tillie-Leblond I, Chanez P, Escamilla R, Court-Fortune I, Perez T, Carré P, Roche N. Clinical COPD phenotypes: a novel approach using principal component and cluster analyses. *European Respiratory Journal* 2010; 36: 531–539.

11. Martínez-Garćia MA, Vendrell M, Girón R, Máiz-Carro L, De La Rosa Carrillo D, De Gracia J, Olveira C. The multiple faces of non-cystic fibrosis bronchiectasis a cluster analysis approach. *Annals of the American Thoracic Society* 2016; 13: 1468–1475.

12. Haldar P, Pavord ID, Shaw DE, Berry MA, Thomas M, Brightling CE, Wardlaw AJ, Green RH. Cluster analysis and clinical asthma phenotypes. *American Journal of Respiratory and Critical Care Medicine* 2008; 178: 218–224.

13. Burgel P-R, Lemonnier L, Dehillotte C, Sykes J, Stanojevic S, Stephenson AL, Paillasseur J-L. Cluster and CART analyses identify large subgroups of adults with cystic fibrosis at low risk of 10-year death. *European Respiratory Journal* 2019; 53: 1801943.

14. Conrad DJ, Bailey BA. Multidimensional clinical phenotyping of an adult cystic fibrosis patient population. *PLoS ONE* 2015; 10: e0122705.

15. Hebestreit H, Hulzebos EHJ, Schneiderman JE, Karila C, Boas SR, Kriemler S, Dwyer T, Sahlberg M, Urquhart DS, Lands LC, Ratjen F, Takken T, Varanistkaya L, Rücker V, Hebestreit A, Usemann J, Radtke T. Cardiopulmonary exercise testing provides additional prognostic information in cystic fibrosis. *American Journal of Respiratory and Critical Care Medicine* 2019; 199: 987–995.

16. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria; 2019.

17. Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K, Studer M, Roudier P, Gonzalex J, Kozlowski K, Schubert E, Murphy K. "Finding Groups in Data": Cluster Analysis Extended Rousseeuw et al. 2019 p. 1–82.

18. Quanjer PH, Stanojevic S, Cole TJ, Baur X, Hall GL, Culver BH, Enright PL, Hankinson JL, Ip MSM, Zheng J, Stocks J, ERS Global Lung Function Initiative. Multi-ethnic reference values for spirometry for the 3-95-yr age range: The global lung function 2012 equations. *European Respiratory Journal* 2012; 40: 1324–1343.

19. Pinheiro J, Bates D, DebRoy S, Sarkar D, authors E, Heisterkamp S, Van Willigen B, R-core. Linear and Nonlinear Mixed Effects Models. 2020.

20. Altman NS. An introduction to kernel and nearest-neighbor nonparametric regression. *American Statistician* 1991; 46: 175–185.

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

22. National Center for Health Statistics. CDC Growth Charts for the United States: Methods and Development. 2000 p. 1–189.

23. Matheson FI, Ingen T van. 2016 Ontario Marginalization Index: User Guide. Toronto ON; 2018 p. 1–19.

24. Hafen GM, Hurst C, Yearwood J, Smith J, Dzalilov Z, Robinson PJ. A new scoring system in Cystic Fibrosis : statistical tools for database analysis – a preliminary report. *BMC Medical Informatics and Decision Making* 2008; 8: 1–11.

25. Moore WC, Meyers DA, Wenzel SE, Teague WG, Li H, Li X, D'Agostino R, Castro M, Curran-Everett D, Fitzpatrick AM, Gaston B, Jarjour NN, Sorkness R, Calhoun WJ, Chung KF, Comhair SAA, Dweik RA, Israel E, Peters SP, Busse WW, Erzurum SC, Bleecker ER. Identification of asthma phenotypes using cluster analysis in the severe asthma research program. *American Journal of Respiratory and Critical Care Medicine* 2010; 181: 315–323.

26. Van Rooden SM, Heiser WJ, Kok JN, Verbaan D, Van Hilten JJ, Marinus J. The identification of Parkinson's disease subtypes using cluster analysis: A systematic review. *Movement Disorders* 2010; 25: 969–978.

27. Nkam L, Lambert J, Latouche A, Bellis G, Burgel PR, Hocine MN. A 3-year prognostic score for adults with cystic fibrosis. *Journal of Cystic Fibrosis* The Authors; 2017; 16: 702–708.

28. Liou TG, Adler FR, FitzSimmons SC, Cahill BC, Hibbs JR, Marshall BC. Predictive 5-Year Survivorship Model of Cystic Fibrosis. *American Journal of Epidemiology* 2001; 153: 345–352.

29. VanDevanter DR, Wagener JS, Pasta DJ, Elkin E, Jacobs JR, Morgan WJ, Konstan MW. Pulmonary Outcome Prediction (POP) Tools for Cystic Fibrosis Patients. *Pediatric Pulmonology* 2010; 45: 1156–1166.

30. McCarthy C, Dimitrov BD, Meurling IJ, Gunaratnam C, McElvaney NG. The CF-ABLE score: A novel clinical prediction rule for prognosis in patients with cystic fibrosis. *Chest* The American College of Chest Physicians; 2013; 143: 1358–1364.

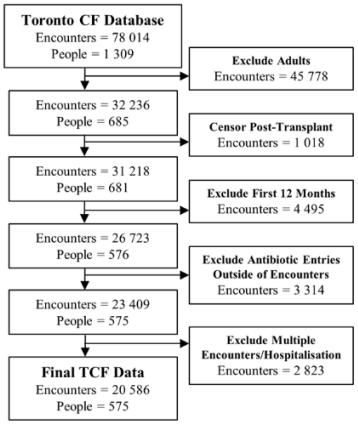
Toronto CF Database
Encounters = 78 014
People = 1 309
Exclude Adults
Encounters = 45 778
Encounters = 32 236
People = 685
Censor Post-Transplant
Encounters = 1 018
Encounters = 31 218
People = 681
Exclude First 12 Months
Encounters = 4 495
Encounters = 26 723
People = 576
Exclude Antibiotic Entries Outside of Encounters
Encounters = 3 314
Encounters = 23 409
People = 575
Exclude Multiple Encounters/Hospitalisation
Encounters = 2 823
Final TCF Data
Encounters = 20 586
People = 575

Figure 1

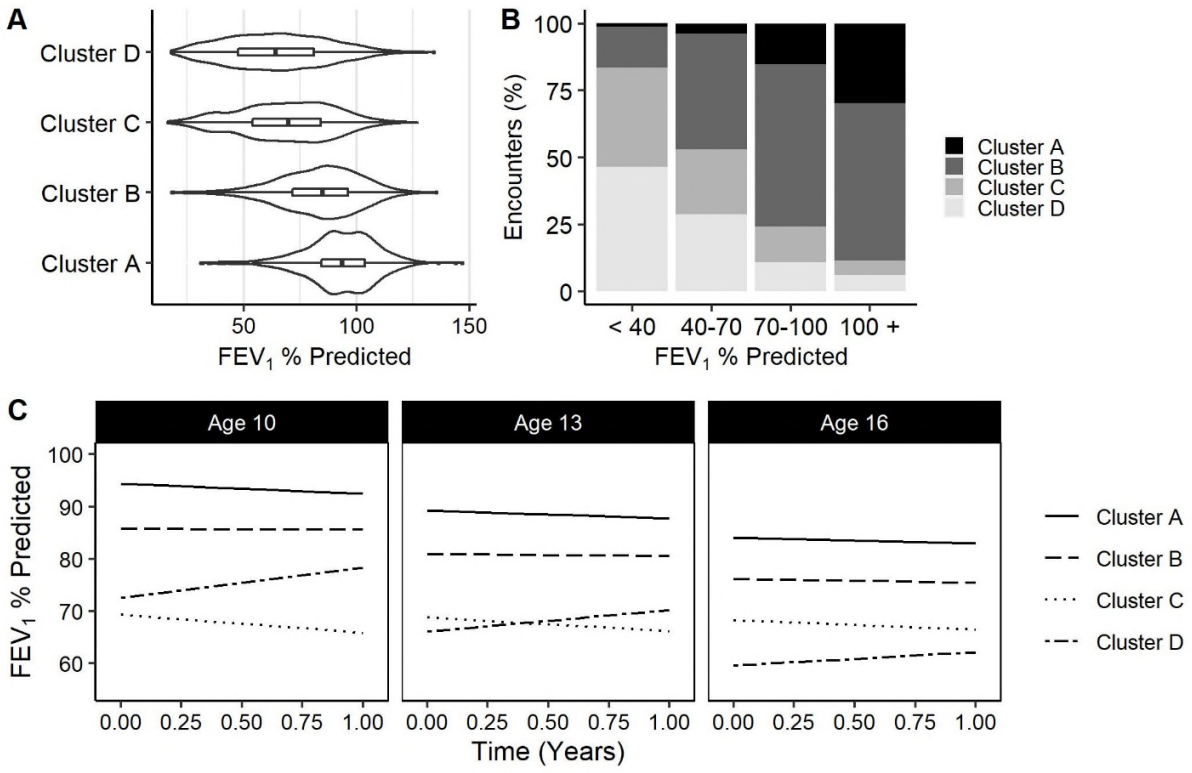
A
Cluster D
Cluster C
Cluster B
Cluster A
50
100
150
FEV1 % Predicted
B
Encounters (%)
100
75
50
25
0
< 40
40-70
70-100
100 +
FEV1 % Predicted
Cluster A
Cluster B
Cluster C
Cluster D
C
Age 10
Age 13
Age 16
FEV1 % Predicted
100
90
80
70
60
0.00
0.25
0.50
0.75
1.00
Time (Years)
Cluster A
Cluster B
Cluster C
Cluster D

Figure 2

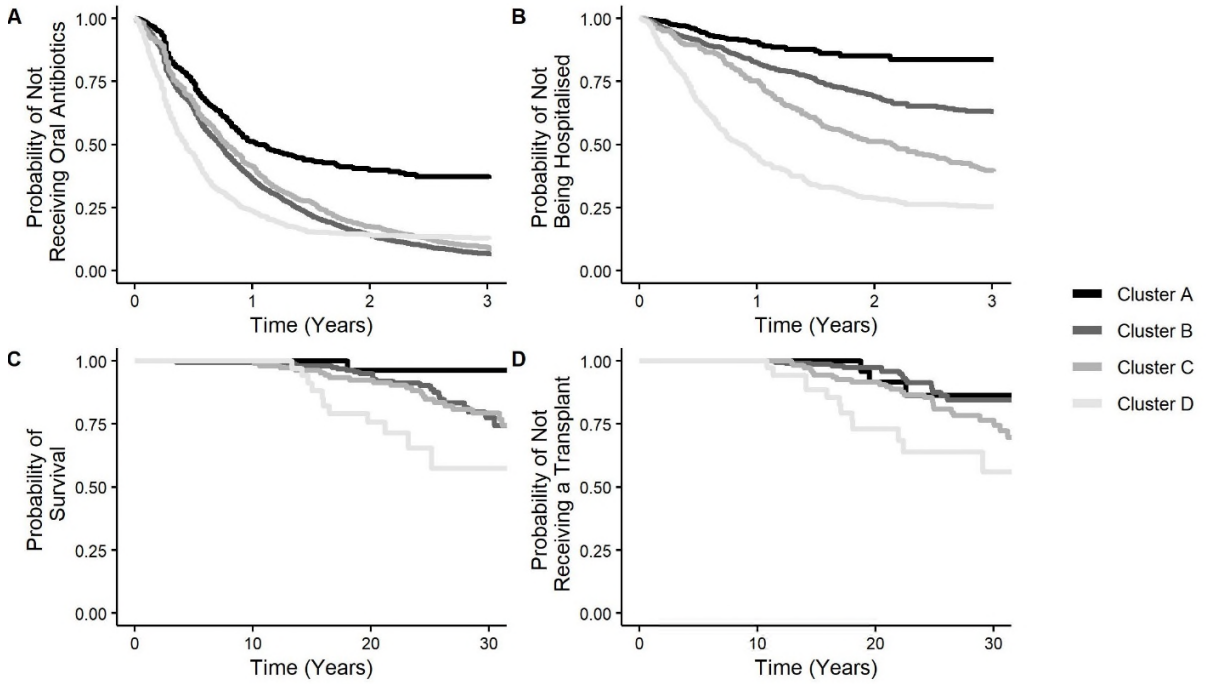
A
Probability of Not Receiving Oral Antibiotics
1.00
0.75
0.50
0.25
0.00
0
1
2
3
Time (Years)
B
Probability of Not Being Hospitalised
1.00
0.75
0.50
0.25
0.00
0
1
2
3
Time (Years)
C
Probability of Survival
1.00
0.75
0.50
0.25
0.00
0
10
20
30
Time (Years)
D
Probability of Not Receiving a Transplant
1.00
0.75
0.50
0.25
0.00
0
10
20
30
Time (Years)
Cluster A
Cluster B
Cluster C
Cluster D

Figure 3

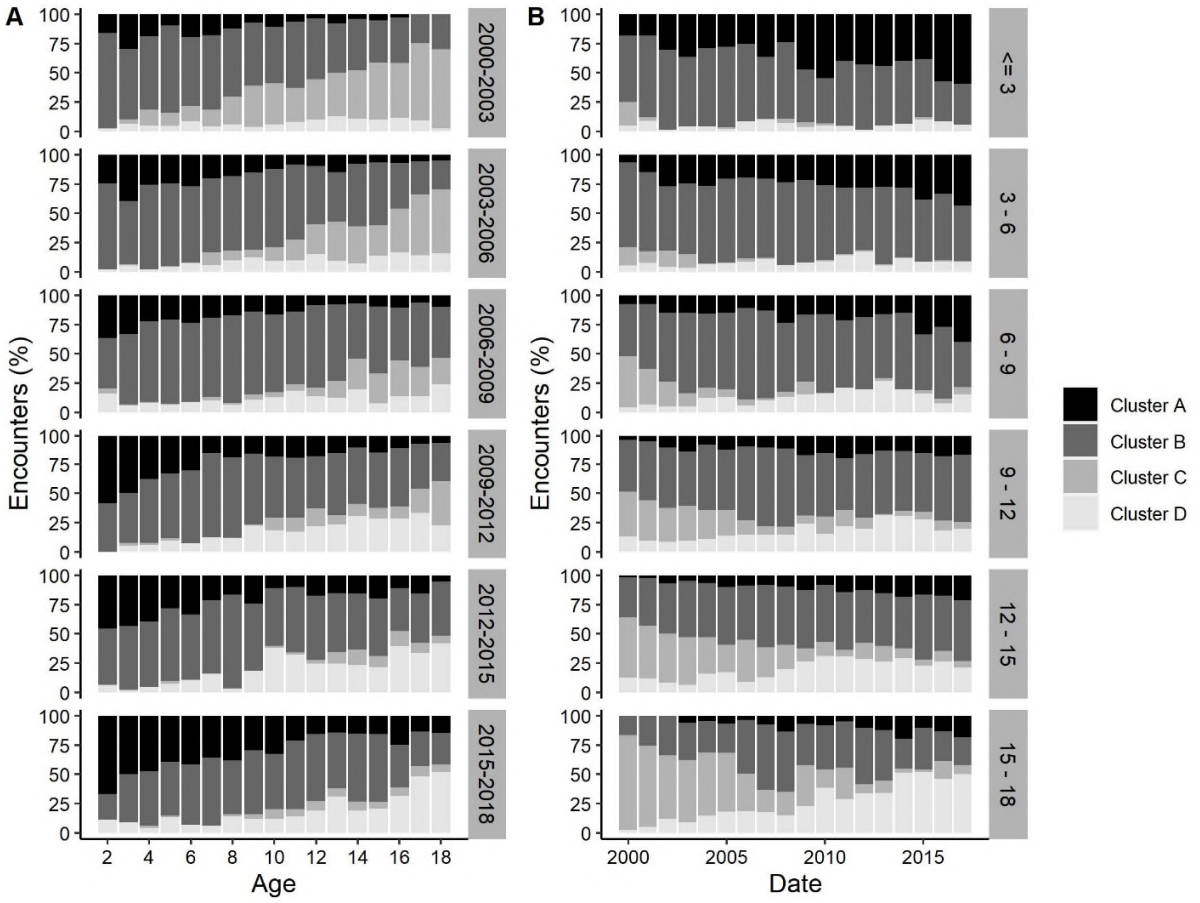
A
B
Encounters (%)
Age
Date
2000-2003
2003-2006
2006-2009
2009-2012
2012-2015
2015-2018
<= 3
3 - 6
6 - 9
9 - 12
12 - 15
15 - 18
Cluster A
Cluster B
Cluster C
Cluster D

Figure 4

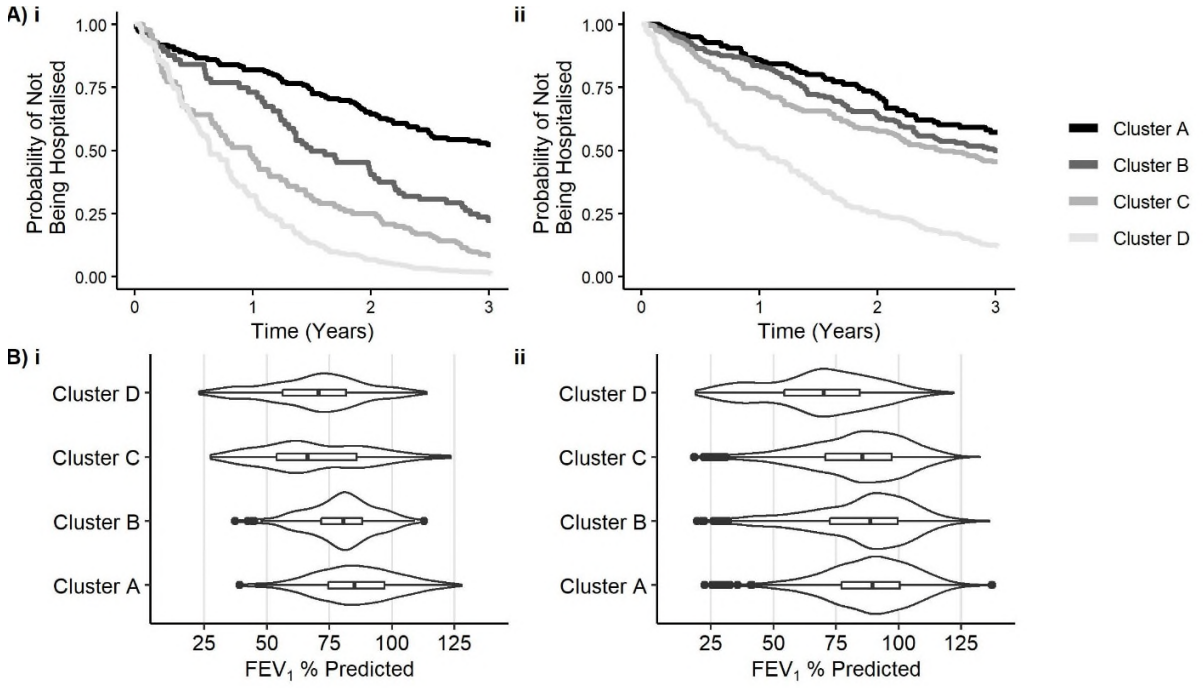
A) i
ii
Probability of Not Being Hospitalised
1.00
0.75
0.50
0.25
0.00
0
1
2
3
Time (Years)
Cluster A
Cluster B
Cluster C
Cluster D
B) i
ii
Cluster D
Cluster C
Cluster B
Cluster A
25
50
75
100
125
FEV1 % Predicted

Figure 5

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).

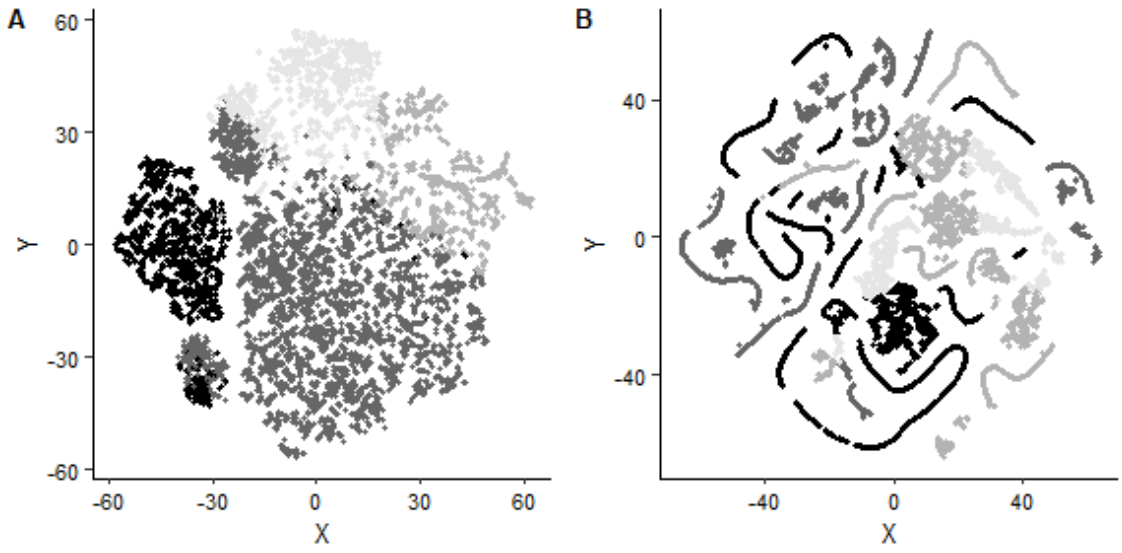

Figure S1. t-SNE plots of A) the optimal model with good cluster distinctions (9 variables: body mass index (BMI), height, hospitalisations in prior year, cough, and previous rates of infection with *Pseudomonas aeruginosa*, *Staphylococcus aureus*, *Stenotrophomonas sp.*, *Haemophilus influenzae*, and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

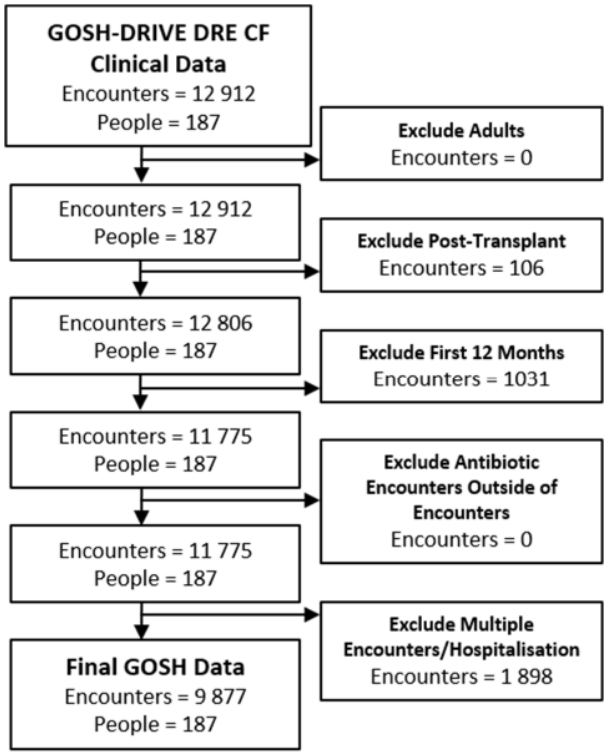

Figure S2. Description of data exclusions for the GOSH Data

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.