Heterogeneity of respiratory disease in children and young adults with sickle cell disease

Alan Lunt, Lucy Mortimer, David Rees, Sue Height, Swee Lay Thein,

Anne Greenough

Supplementary file

Lung function testing

The subjects had been recruited into three previously published prospective studies.[1-3] between February 2010 and April 2015. Lung function testing was conducted in the Amanda Smith Unit at King's College Hospital NHS Foundation Trust. No participant underwent testing within two weeks of an upper respiratory tract infection or within a month of suffering a vaso-occlusive crisis. A history was taken of past and current respiratory symptoms and medication for respiratory problems. Standing height was measured using a wall-mounted stadiometer (Holtain ltd, Crymych, Dyfed, UK), and weight using electronic weighing scales (Seca ltd, Birmingham, UK). Subjects were assessed while wearing a nose-clip and breathing through a mouthpiece. Respiratory system resistance at a frequency of 5Hz (R_5) was measured during a 90 second period of tidal breathing using impulse oscillometry (IOS, Jaeger Masterscreen IOS, Carefusion Ltd, Basingstoke UK). The results were expressed as the percent predicted for height.[4] The mean of two measurements within 10% of each other was reported. IOS was performed before spirometry and plethysmography, that is prior to any changes in bronchial smooth muscle tone caused by deep inspiration during those measurements. Spirometry and lung volume measurements were performed using a pneumotachograph based system (Jaeger Masterscreen PFT,

Carefusion Ltd, Basingstoke UK). Forced expiratory volume in one second (FEV_1), vital capacity (VC), mean maximum expiratory flow (FEF_{25-75}), total lung capacity (TLC), residual volume (RV), transfer factor for carbon monoxide (DLCO), and transfer coefficient (K_{CO}) were assessed and the results expressed as the percent predicted for height, sex, age and race where appropriate.[5, 6] DLCO and K_{CO} were corrected for haemoglobin concentration. Spirometry was repeated after administration of 400mcg salbutamol via MDI and spacer and an increase in FEV_1 of twelve percent or greater was considered a positive response.[7] Ethnic-specific reference equations were not available for static lung volumes or gas transfer; therefore the predicted values were adjusted for people of African descent using appropriate correction factors (-12% for adults and -6% for children).[7, 8] Patients were diagnosed with a restrictive abnormality if their TLC was less than the lower limit of normal (LLN) with a normal FEV_1:VC, an obstructive abnormality if their FEV_1:VC was less than LLN and a mixed pattern if both their TLC and FEV_1:VC were less than the LLN.[7]

Analysis

Inclusion of a large number of variables for small datasets may degrade the final classification produced by cluster analysis [9, 10], therefore, the smallest number of biomarkers necessary to adequately characterise previously described lung function or haematological abnormalities, or which relate to clinical severity were used Functionally redundant variables were not included (for example those which did not give additional information to characterise the lung function abnormalities). This included VC, RV, FRC, FEF_{25-75} and KCO when FEV/VC and TLC were available.

Tree-based methods select variables based on their capacity to discriminate between categories of the response variable (in this case cluster number) and produce a simple decision tree which can be applied to data from a new subject to estimate, with optimal

reliability, to which cluster the subject is most likely to belong. Importantly, with a sufficiently reliable tree model, knowledge of all the variables used in the clustering is not required, and thus these methods may provide an easily implementable method of stratification based on readily available clinical measurements.

Clustering was performed for solutions comprising two to ten clusters and the solution with the highest CritCF was considered optimal. All results were standardized prior to clustering using a nonparametric standardisation procedure (the median value for the column variable was subtracted from each data point and the result was divided by the median absolute deviation of the column variable). Multiple imputation using chained equations was used to assess the impact of missing values on the clustering solution.[11] Fifty imputed datasets were generated using all variables included in the clustering.[12] The clustering procedure described above was then performed on each of the imputed datasets and the CritCF was used to select an optimal clustering solution for each imputed dataset. The distribution of clustering solutions as a function of imputation was then examined and the clustering solution that occurred most frequently in the fifty imputed datasets was used. The clustering was visualised using a discriminant-coordinates biplot, generated by canonical variate analysis, which projects multidimensional data into a lower dimensional space while preserving as much information as possible, to provide an easily interpretable two-dimensional representation of cluster separation and density. Ninety-five percent confidence regions were also derived for the clusters.[13] Variables were then compared across clusters using Kruksal-Wallis tests with post-hoc multiplicity-adjusted pairwise comparisons. Cluster profiles were presented graphically using a radial plot. Radial plots provide a method of displaying multiple quantitative variables on a single polar grid, where the length of each ‘spoke’ is proportional to the magnitude of the standardised variable. A conditional inference tree analysis was used to derive a stratification algorithm to predict cluster membership based on minimal subset of input variables (R package “party”, version 1.1-

25).[14] Tree-based methods select variables based on their capacity to discriminate between categories of the response variable (in this case cluster number) and produce a simple decision tree which can be applied to data from a new subject to estimate, with optimal reliability, to which cluster the subject is most likely to belong. Importantly, with a sufficiently reliable tree model, knowledge of all the variables used in the clustering is not required, and thus these methods may provide an easily implementable method of stratification based on readily available clinical measurements. The dataset was randomly partitioned into a "training set", comprising seventy-five percent of the cohort used in the cluster analysis, with the remaining twenty-five percent forming the "validation set". The training set was used to derive a conditional inference tree model, which was then tested using the unseen data from the validation set to assess the predictive accuracy of the model when classifying new data.

RESULTS

Data for one or more haematological variables were missing in 24 patients (21%), accounting for a total of 6.5% of the dataset (i.e. 6.5% of a total possible dataset of 11 measurements from each of 114 patients). Missing data rates were similar across the clusters (p=0.580). Twenty-one patients (18%) had an obstructive, twenty-three (20%) a restrictive and ten (9%) a mixed ventilatory defect. Thirty-eight patients (33%) had a history of ACS and nine (7%) had a significant response to bronchodilator. Fifteen (13.1%) were taking hydroxyurea and twenty-six (22.8%) were receiving regular blood transfusions. Data on non-elective admissions and ACS episodes were available for a period of median 4.2 (2.3 – 6.8) years following testing in 103 patients. The frequency of hospital admissions (p=0.025), but not ACS (p=0.821), was significantly different between clusters. On post-hoc testing, cluster three had a greater frequency of hospital admissions than cluster one and cluster two (0.69 (0.0 – 6.1) events/year versus 0.25 (0.0 – 1.50) events/year and 0.38 (0.0 –

3.02) events/year, respectively, both $p<0.05$). Those results were unchanged after adjusting for differences in hydroxyurea use and chronic transfusion.

DISCUSSION

This study has strengths and some limitations. A strength was the use of a well-characterised cohort of patients with SCD, spanning a wide age range. We also assessed the impact of missing data on the clustering solution. Our data were from a single centre, which obviates the potential effects of site-specific variations in equipment and measurement protocols, but may limit the generalisability of the results, which therefore require evaluation in an external validation cohort. Longitudinal assessment of subjects would be required to evaluate the stability of the phenotypic groups over time. Additionally, our data included relatively low (13%) and high (22.8%) proportions of patients on hydroxyurea and chronic transfusion therapy, respectively. This too might impact on the generalisability of our clustering solution. A further limitation might appear to be that this was a secondary analysis of existing data and complete haematological data were not available for all patients. Our multiple imputation analysis, however, suggested that missing data did not have a significant impact on the clustering obtained. Our use of haematological data obtained for routine clinical use did restrict the variables we were able to include in the analysis. It might have been informative to add HbF as high concentrations are known to exert a protective effect against HbS polymerisation and may lead to a milder phenotype [15], but these data were not available in a large enough number of patients. We expressed the lung function results as a percentage of predicted values rather than z-scores, but lung function abnormalities were defined using the lower limit of normal as recommended by current guidelines. Whilst z-scores are used in the paediatric literature, their use has not been adopted widely in studies of adult lung function. Since our cohort contained a substantial proportion of adult patients, we expressed the lung function results as a percentage of predicted which would

be understood to both audiences. We used ethnic-specific reference values for spirometric results, but these are not available for static lung volumes and IOS. Static lung volume and impulse oscillometry results were related to reference ranges derived from Caucasian subjects, but the two groups were matched for ethnic origin and the same reference ranges were used in all patients, thus the comparisons between the clusters were valid. We used conditional inference tree analysis to derive our stratification algorithm based on a reduced set of variables; tree based methods may attribute more weight to larger clusters, which may be problematic where small clusters are thought to be of greater importance. In these circumstances, the use of group size weighted methods of classification may be helpful. In our analysis, however, poorer outcomes were seen only in cluster three (n=45), so that limitation is unlikely to have substantially influenced interpretation of our results. We have identified three distinct phenotypes of children and young adults with sickle cell disease. The clusters were associated with different patterns of lung function impairment and haematological variables and hence may reflect different disease processes. The classification of SCD patients by phenotypes may help to provide better management.

REFERENCES

1. Wedderburn CJ, Rees D, Height S, et al. Airways obstruction and pulmonary capillary blood volume in children with sickle cell disease. *Pediatr Pulmonol* 2014;**49**:716-22.
2. Lunt A, Ahmed N, Rafferty GF, et al. Airway and alveolar nitric oxide production, lung function and pulmonary blood flow in sickle cell disease. *Pediatr Res* 2016;**79**:313-7.
3. Lunt A, McGhee E, Sylvester K, et al. Longitudinal assessment of lung function in children with sickle cell disease. *Pediatr Pulmonol* 2016;**51**:717-23.
4. Nowowiejska B, Tomolak W, Radlinski J, et al. Transient reference values for impulse oscillometry for children aged 3–18 years. *Pediatr Pulmonol* 2008;**43**:1193-97.
5. Quanjer PH, Stanojevic S, Cole TJ, et al. Multi-ethnic reference values for spirometry for the 3-95 year age range: the global lung function 2012 equations. *Eur Respir J* 2012;**40**:1324-43.
6. Rosenthal M, Cramer D, Bain SH, et al. Lung function in white children aged 4 to 19 years: II--Single breath analysis and plethysmography. *Thorax* 1993;**48**:803-8.
7. Pellegrino R, Viegi G, Brusasco V, et al. Interpretative strategies for lung function tests. *Eur Respir J* 2005;**26**:948-68.
8. Kirkby J, Bonner R, Lum S, et al. Interpretation of pediatric lung function: impact of ethnicity. *Pediatr Pulmonol* 2013;**48**:20-6.
9. Garcia-Aymerich J, Gómez FP, Benet M, et al. Identification and prospective validation of clinically relevant chronic obstructive pulmonary disease (COPD) subtypes. *Thorax* 2011;**66**:430-7.
10. Wang Y, Miller DJ, Clarke R. Approaches to working in high-dimensional data spaces: gene expression microarrays. *Br J Cancer* 2008;**98**:1023-8.

11. Basagana X, Barrera-Gómez J, Benet M, et al. A framework for multiple imputation in cluster analysis*. Am J Epidemiol* 2013;**177**:718-25.
12. van Buuren S, Groothuis-Oudshoorn K. Mice: Multivariate Imputation by Chained Equations in R. *J Stat Software* 2011;**45**:1-67.
13. Gardner S, le Roux RJ. Extensions of biplot methodology to discriminant analysis. *J Classification* 2005;**22**:59-86.
14. Strobl C, Malley J, Tutz G. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychol Methods* 2009;**14**:323-48.
15. Franco RS, Yasin Z, Palascak MB, et al. The effect of fetal hemoglobin on the survival characteristics of sickle cells. *Blood* 2006;**108**:1073-6.

Table A: Characteristics of the entire cohort

Results are expressed as median (IQR).

BDR: bronchodilator reversibility

*Lung function tests are expressed as the percentage predicted for age and/or height.

Age (yrs)	14.5 (10.6 – 17.9)
FEV_1:VC*	93.7 (88.5 – 100.8)
R_{rs5}*	134.4 (113.0 – 161.2)
TLC*	87.1 (80.4 – 97.1)
D_LCO*	87.4 (78.4 – 96.6)
PCBV (ml/l)	24.0 (20.9 – 29.6)
[Hb] (g/dl)	8.9 (7.9 – 10.1)
S_pO_2 (%)	96.2 (94.0 – 99.0)
LDH (IU/L)	495 (387.8 – 649.2)
WCC ($x10^9$/L)	10.1 (6.8 – 12.0)
Reticulocytes (%)	9.8 (7.0 – 13.5)
ACS ever	33.3%
Lung function abnormalities:	
Obstructive n (%)	21 (18%)
Restrictive n (%)	23 (20%)
Mixed n (%)	10 (9%)
BDR%	7%

FIGURE LEGENDS

Figure S1:

A: Box plot of the between-imputation distribution of CritCF by number of clusters.

B: Selection frequency of clustering solution over multiple imputations by number of clusters. Summary data for the three clusters are given in Table 1

Figure S2:

Discriminant projection plot of clustering solution. The inset shows the projection basis vectors: arrows indicate the direction of increase for each variable and the line length reflects the extent to which each variable contributes to the point coordinates. Shaded areas are the 95% confidence regions for each cluster assuming eliptical clusters. Subjects in the validation set who were misclassified by the conditional inference tree are indicated by circled points.

Figure S3:

Radial plots showing physiological profiles for the three clusters. Data are standardized (expressed as z-scores referenced to the whole cohort) and the points represent medians. The dashed circle is the whole cohort average.