Predicting healthcare outcomes in prematurely born infants using cluster analysis

Victoria MacBean PhD[1], Alan Lunt PhD[2,3], Simon B Drysdale PhD[4], Muska N Yarzi BSc[5], Gerrard F Rafferty PhD[1], Anne Greenough MD (Camb)[2,36]

[1] Faculty of Life Sciences and Medicine, King's College London

[2]MRC & Asthma UK Centre in Allergic Mechanisms of Asthma, King's College London, United Kingdom

[3] Department of Women and Children's Health, School of Life Course Sciences, Faculty of Life Sciences and Medicine, King's College London, United Kingdom

[4] Oxford Vaccine Group, Department of Paediatrics, University of Oxford, United Kingdom

[5] Cellular and Molecular Medicine, University of Bristol, United Kingdom

[6] NIHR Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust and King's College London, United Kingdom

Financial support: Dr Vicky MacBean was supported by the King's Health Partners Challenge Fund. The research was supported by the National Institute for Health Research (NIHR) Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust and King's College London. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

Address for correspondence and reprint requests author: Anne Greenough, Neonatal Intensive Care Unit, 4th Floor Golden Jubilee Wing, King's College Hospital, Denmark Hill, London, SE5 9RS, United Kingdom. Tel: 0203 299 3037; fax: 0203 299 8284 Email: anne.greenough@kcl.ac.uk

Key words: prematurity, healthcare utilisation, prediction, respiratory viruses

Running head: Cluster analysis and outcomes after preterm birth

ABSTRACT

Aims: Prematurely born infants are at high risk of respiratory morbidity following neonatal unit discharge, though prediction of outcomes is challenging. We have tested the hypothesis that cluster analysis would identify discrete groups of prematurely born infants with differing respiratory outcomes during infancy.

Methods: 168 infants (median (IQR) gestational age 33 (31 – 34) weeks) were recruited in the neonatal period from consecutive births in a tertiary neonatal unit. The baseline characteristics of the infants were used to classify them into hierarchical agglomerative clusters. Rates of viral lower respiratory tract infections (LRTIs) were recorded for 151 infants in the first year after birth.

Results: Infants could be classified according to birth weight and duration of neonatal invasive mechanical ventilation (MV) into three clusters. Cluster one (MV ≤5 days) had few LRTIs. Clusters two and three (both MV ≥6 days, but BW ≥ or <882g respectively) had significantly higher LRTI rates. Cluster two had a higher proportion of infants experiencing respiratory syncytial virus LRTIs (p=0.01) and cluster three a higher proportion of rhinovirus LRTIs (p<0.001)

Conclusions: Readily available clinical data allowed classification of prematurely born infants into one of three distinct groups with differing subsequent respiratory morbidity in infancy.

INTRODUCTION

Infants born prematurely are at higher risk of lower respiratory tract infections (LRTIs) in infancy, which may in themselves be associated with poorer respiratory health. We have shown that infants born prematurely have high rates of respiratory syncytial virus (RSV) LRTIs and subsequent respiratory morbidity in the first two years after birth [1; 2]. LRTIs related to other viral pathogens, including human rhinovirus (RV), may also be associated with poorer infant outcomes in those born prematurely [3; 4]. Although the risk is elevated above that for term-born infants, LRTI rates of between 46% and 58% have been reported, with fewer than five per cent of infants requiring hospitalisation [2; 5].

Prematurely born children are also at higher risk of respiratory morbidity in later childhood than term-born infants. The severity of any respiratory disease and the healthcare burden this poses, however, varies substantially amongst children born prematurely. Cohort studies have shown that while group average values for pulmonary function in children born prematurely were lower than those of matched term-born controls, many children were within the normal range [6; 7].

The heterogeneity in those adverse consequences of preterm birth highlights the importance of being better able to prognosticate both to support clinicians in counselling parents in the neonatal period and to guide researchers in identifying high risk groups to target for interventional studies. Prophylactic agents against respiratory viral infections are becoming increasingly available, but are expensive, hence identifying infants at highest risk for adverse outcomes is desirable. A number of studies have identified risk factors for poorer outcomes following preterm birth. We have demonstrated that certain single nucleotide polymorphisms were associated with a higher risk of RSV LRTIs, greater respiratory morbidity in the first year after birth and poorer pulmonary function at one year of age [8]. While undertaking genetic profiling is becoming both easier and less costly, such data are still not routinely available within clinical practice. Other studies have identified specific risk factors associated with poorer outcomes, including duration of supplemental oxygen requirement, gestational age, sex, intrauterine growth restriction, race and socio-demographic factors [9; 10]. The data from such studies are of great epidemiological value, though are of less benefit to clinicians attempting to determine the prognosis of an individual infant. Additionally, those studies have often focussed on the clinical course within the first year after birth and have not provided data on longer-term outcomes.

The aim of the current study was to examine whether groups of phenotypically distinct infants (clusters) could be identified based on the detailed clinical and physiological data collected at baseline in a large prospectively recruited cohort of prematurely born infants. Further aims were to determine whether a simplified set of routinely available clinical characteristics could be used to reliably classify infants into identified groups. Furthermore, we aimed to determine whether such groups had different risk profiles for LRTIs in infancy, specifically those caused by RSV and RV. An exploratory analysis investigated whether there were any differences in respiratory outcomes between groups in a subset of the children who were followed up to school age.

METHODS

Study subjects

A cohort of 168 prematurely born infants were assessed who had been enrolled into a study investigating the association between premorbid lung function abnormalities and risk of viral LRTIs in prematurely born infants [11]. They consisted of consecutive births outside the RSV season (defined as 1st October to 1st March) at King's College Hospital NHS Foundation Trust (KCH) whose parents gave informed written consent for their infant's participation in the study. Clinical variables were recorded including birth weight, gestational age, days of invasive ventilation and duration of supplemental oxygen requirement. Assessment of pulmonary function was made at 36 weeks' postmenstrual age (PMA), and comprised measurement of functional residual capacity by helium dilution (FRCHe) and respiratory system compliance (Crs) and resistance (Rrs). Full details of the methodology for infant pulmonary function testing have been given elsewhere [12]. All values were reported as standardised residuals ('z scores') relative to published reference data for FRCHe [13] and Crs and Rrs [14].

Throughout the first year after birth, parents were asked to contact the research team whenever the infant displayed signs consistent with an LRTI. Parents also received a fortnightly telephone call as a reminder. If an LRTI was reported, a researcher visited the infant at home or in hospital to obtain a nasopharyngeal aspirate (NPA). NPA samples were tested for thirteen respiratory viruses (rhinovirus, human metapneumovirus, influenza A and B, parainfluenza 1-3, RSV A and B, enterovirus and parechovirus *via* real-time reverse transcriptase polymerase chain reaction (PCR), and human bocavirus and adenovirus *via* real-time PCR).

At five to seven years of age all previous participants were contacted and invited to return for follow-up measurements [11]. Parents gave further informed written consent for the child to take part in the follow up study. Pulmonary function was assessed by parasternal intercostal electromyography, which provides a non-volitional measure of respiratory effort. We have previously shown this technique detects changes induced by imposition or removal of additional respiratory load in children [15; 16]. Healthcare utilisation was recorded and cost of care calculated. The follow up study was granted ethical approval from the National Research Ethics Service Committee West Midlands – Coventry & Warwickshire (reference 15/WM/0117).

Assessment of pulmonary function

The parasternal intercostal electromyogram (EMGpara) was undertaken in accordance with published work [15]. Briefly, EMGpara was recorded during ten minutes of tidal breathing using surface electrodes placed over the second intercostal space, directly adjacent to the sternum. The mean peak EMGpara per breath over the final stable minute of recording was reported and expressed as z scores based on previously reported data [15]. Further details are given in the online supplementary material.

Healthcare utilization and health related cost of care

Participants' General Practitioner (GP) records were inspected to identify any hospital admissions, emergency department visits, hospital outpatient appointments, other contacts with health professionals, GP attendances and all medication prescriptions.

In the United Kingdom's National Health Service, primary care providers (General Practitioners) hold a comprehensive record of all healthcare contacts as they act as 'gatekeepers' for all healthcare access (including secondary and specialist care by any medical, nursing or allied health professional). Any visits for routine immunisations or health screening were not included in the analysis as these were deemed usual care. GP and hospital costs were calculated using the National Health Service (NHS) reference costing scheme [17] and medication costs using the NHS indicative costs listed within the British National Formulary [18]. Each healthcare contact or prescription was classed as respiratory-related or non respiratory-related. All healthcare costs were divided by the number of years of follow-up and expressed as UK pounds (£) per year for medication costs, hospital costs (inpatient stays, outpatient appointments and emergency department visits) and overall health related costs of care, with GP attendances expressed as number of visits per year.

Statistical analysis

The baseline characteristics of the cohort of 168 infants were used to classify them into clusters using hierarchical agglomerative clustering with the Euclidean distance metric, using Ward's method. Variables included in the clustering were gestational age, birth weight, duration of mechanical ventilation, duration of supplemental oxygen and z-scores for FRC, Rrs and Crs. Gestational age, birth weight, duration of invasive mechanical ventilation and duration of supplemental oxygen 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). The optimal number of clusters was selected using the majority-rule system as implemented in the R package ‘NbClust’. NbClust implements thirty cluster quality indices; each index provides an estimate of the optimal number of clusters in the dataset and the cluster number most frequently selected was used for the final classification [19].

The clustering was visualised using a discriminant-coordinates plot, 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 [20]. A conditional inference tree analysis was used to derive a stratification algorithm to predict cluster membership based on a minimal subset of input variables (R package “party”, version 1.1-25) [21]. 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 a simple-to-use tool for 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.

Clusters were then compared using the non-parametric Kruskal-Wallis test (as the smaller clusters either contained too few individuals to allow for normality testing or demonstrated non-Gaussian distributions) *post hoc* testing using the Mann-Whitney test for differences between groups was undertaken.

RESULTS

A three-cluster solution was identified as optimal for the full cohort of 168 infants. All baseline variables were significantly different across clusters, with the exception of sex (Table 1).

Infants could be classified with 100% accuracy into the clusters, as defined from the larger dataset, using a conditional inference tree (Figure 1). All infants with a duration of ventilation of five days or fewer fitted cluster one. Infants with six or more days of invasive ventilation were classified into cluster two or cluster three according to a birth weight of ≥ or < 882g respectively.

First year after birth

A total of 151 infants were followed up to one year of age. Infants in cluster one were significantly less likely to have experienced any viral LRTI (p=0.02), with a significantly higher proportion of infants in cluster three experiencing an LRTI (p=0.02) (Table 2). Infants in cluster one had a significantly higher proportion of infants with no LRTI (p=0.02) and a lower proportion of infants with RSV LRTI (p=0.04). Cluster two had a significantly increased proportion of infants experiencing RSV LRTI (p=0.01). Infants in cluster three were significantly more likely to have experienced RV LRTI (p<0.001) and less likely to have had no LRTI (p=0.02) (Table 2).

School age follow up

At school age follow-up of 56 children, the clusters comprised 45 children in cluster one, six in cluster two and five in cluster three.

EMGpara was significantly different (p= 0.012) between groups. *Post hoc* testing showed that median (IQR) EMGpara was significantly higher (that is worse) in clusters two (1.93 (1.52 to 2.01) z scores, p=0.041) and three (2.23 (1.67 to 2.86) z scores, p=0.014) compared to cluster one (0.99 (0.41 to 1.57) z scores).

Healthcare utilisation costs, with the exception of non-respiratory related GP attendances and medication costs, were significantly different across clusters (Table 3). Total non-respiratory related healthcare costs (p=0.002) and non-respiratory related hospital costs (p=0.001) were significantly higher in cluster two than cluster one. Total respiratory-related (p=0.001) and non-respiratory related healthcare costs (p=0.039), respiratory-related GP attendances (p=0.004), respiratory medication costs (p=0.011) and both respiratory (p=0.046) and non-respiratory related (p=0.036) hospital costs were significantly higher in cluster three than cluster one.

DISCUSSION

This study has used multidimensional phenotyping based on cluster analysis to detect and explore distinct groups within an unselected cohort of prematurely-born infants and has demonstrated that the respiratory outcomes in infancy differed significantly across these groups. Furthermore, exploratory analysis of data acquired at school age suggested that the differences persisted into childhood. The success with which infants could be classified into the three clusters using readily available clinical characteristics suggests that this model could be of value to clinicians seeking to determine prognosis at the point of care in the neonatal unit.

Cluster one comprised moderately and late preterm-born infants requiring minimal or no respiratory support in the perinatal period and with generally normal pulmonary function at 36 weeks' gestational age. These infants went on to have few LRTIs in the first year after birth, specifically with a lower likelihood of RSV LRTI. Pulmonary function, quantified by EMGpara, was within the normal range and healthcare costs over the follow-up period were also low. Cluster two were generally very preterm, very low birth weight infants requiring respiratory support in the neonatal period. These infants' pulmonary function at 36 weeks was somewhat heterogeneous, with values ranging from normal to moderate impairment. In the first year after birth this group had a higher likelihood of RSV LRTI. Pulmonary function at school age appeared to be abnormal, indicated by raised EMGpara. The only significant differences in healthcare costs between this cluster and cluster one were those in non-respiratory related categories, with non-respiratory hospital costs and overall non-respiratory costs being higher. Cluster three, while small, was very well-differentiated and consisted of infants at high risk for poor health outcomes. The infants were born very preterm and had extremely low birth weights (less than 1000 grams), required extensive respiratory support after birth and had significantly impaired pulmonary function at 36 weeks of gestation. In the first year after birth the infants had a much higher likelihood of experiencing LRTIs, with a significantly larger proportion developing RV LRTIs. At school age, they appeared to have raised EMGpara and high healthcare utilisation costs for respiratory problems in both primary and secondary care, as well as high non-respiratory related costs.

The outcomes of all three clusters are of clinical as well as research interest. An infant classified into cluster one could be expected to experience very little ill health in their first five to seven years and to have essentially normal lung function at school age. Cluster three in contrast represent a high-risk group for respiratory infections in infancy and a group in whom high healthcare utilisation could be anticipated. Stratifying delivery of clinical services in accordance with these cluster classifications may allow resources to be directed to those children with greatest need. Similarly, the information we present here may support sample size calculations for future research studies wishing to target at risk groups for follow up or novel interventions.

The added value of the current study to previously published data examining risk factors for respiratory morbidity following premature birth [7; 9; 10] is the potential for direct application to individual infants. Studies often present data in terms of relative risk, whereas the decision metric provided here allows prediction for an infant using only their birth weight and duration of mechanical ventilation. The tool can, therefore, be applied to infants in the immediate neonatal period.

Cluster analysis has been previously used to demonstrate an increased incidence of LRTIs in specific geographic regions [22]. It has also been used to identify distinct severe bronchiolitis profiles [23]. Our study adds to literature demonstrating using cluster analysis could be used in prematurely born infants to predict the likelihood of early life and school age morbidity. Future studies should investigate our findings in larger groups of infants

These results also support the hypothesis that RV LRTI in infancy is a marker of underlying susceptibility to longer-term respiratory morbidity, rather than being a causative factor [24; 25]. Those studies were conducted in term-born cohorts and the current study adds to the literature demonstrating an at-risk group (cluster three) within a cohort of prematurely-born infants.

The strengths of this study include the relatively large cohort recruited, the detailed information regarding respiratory infections in the first year after birth and the length of the follow-up period. A weakness of the study is the number of children in whom school age data could be obtained, meaning that the inferences about longer term outcomes between clusters are somewhat less robust than the prediction of LRTI risk. We recognise the possibility of error introduced by the small numbers of individuals in clusters two and three at school age, but have shown a number of significant differences between our three clusters at school age. The main reason for non-participation was failure to respond to the invitation letters. All surviving members of the original cohort were invited to participate, with a number of reminder invitations sent over the one-year period of recruitment. The overall retention rate of 33%, however, was similar to that reported in other studies [26; 27]. The retention rates within clusters were 32%, 30% and 63% for clusters one, two and three respectively. It was fortunate that attrition was lower in the smallest of the three groups such that this group could be retained within the analysis. The cohort of children whom we were able to follow up were representative of the overall group, with the exception of a significantly lower birth weight and longer duration of supplemental oxygen in those children followed up in cluster two. Our focus was on respiratory outcomes and we did not quantify other potential complications of prematurity such as neurodevelopmental abnormalities. Given the apparent higher non-respiratory healthcare costs within clusters two and three, it is possible that such consequences of premature birth were more pronounced within these two groups.

In conclusion, our study provides evidence that readily available clinical variables can be used to effectively classify prematurely born infants into distinct groups. Those groups were at varying degrees of risk for viral respiratory tract infections, and our data suggest that they had differing respiratory outcomes at school age. These data can be applied on an individual basis and, therefore, may be of value to clinicians counselling parents as to the possible outcomes of preterm birth as well as researchers planning interventional studies.

ACKNOWLEDGEMENTS

Competing interest: Professor Greenough has held grants from various manufacturers (Abbot Laboratories, MedImmune) and ventilator manufacturers (SLE). Professor Greenough has received honoraria for giving lectures and advising various manufacturers (Abbot Laboratories, MedImmune) and ventilator manufacturers (SLE). Professor Greenough is currently receiving a non-conditional educational grant from SLE.

Contributor statement: AG, VMacB and GFR designed the study; VMacB, SDB, MNY collected the data; AL, VMacB and AG analysed the data. All authors were involved in producing the manuscript and approved the final version.

REFERENCES

1. Broughton S, Roberts A, Fox G, Pollina E, Zuckerman M, Chaudhry S, Greenough A. 2005. Prospective study of healthcare utilisation and respiratory morbidity due to RSV infection in prematurely born infants. Thorax. 60(12):1039-1044
2. Drysdale S, Alcazar-Paris M, Wilson T, Smith M, Zuckerman M, Peacock J, Johnston S, Greenough A. 2015. Viral lower respiratory tract infections and preterm infants' healthcare utilisation. Eur J Pediatr. 174(2):209-215
3. Drysdale S, Alcazar M, Wilson T, Smith M, Zuckerman M, Lauinger I, Tong C, Broughton S, Rafferty G, Johnston S et al. 2014. Respiratory outcome of prematurely born infants following human rhinovirus A and C infections. Eur J Pediatr. 173(7):913-919.
4. Drysdale S, Alcazar-Paris M, Wilson T, Smith M, Zuckerman M, Broughton S, Rafferty G, Peacock J, Johnston S, Greenough A. 2013. Rhinovirus infection and healthcare utilisation in prematurely born infants. Eur Respir J. 42(4):1029-1036.
5. Drysdale S, Lo J, Prendergast M, Alcazar M, Wilson T, Zuckerman M, Smith M, Broughton S, Rafferty G, Peacock J et al. 2014. Lung function of preterm infants before and after viral infections. Eur J Pediatr. 173(11):1497-1504
6. Fawke J, Lum S, Kirkby J, Hennessy E, Marlow N, Rowell V, Thomas S, Stocks J. 2010. Lung function and respiratory symptoms at 11 years in children born extremely preterm. Am J Respir Crit Care Med. 182(2):237-245.
7. Zivanovic S, Peacock J, Alcazar-Paris M, Lo JW, Lunt A, Marlow N, Calvert S, Greenough A. 2014. Late outcomes of a randomized trial of high-frequency oscillation in neonates. N Engl J Med. 370(12):1121-1130

8. Drysdale S, Prendergast M, Alcazar M, Wilson T, Smith M, Zuckerman M, Broughton S, Rafferty G, Johnston S, Hodemaekers H et al. 2014. Genetic predisposition of RSV infection-related respiratory morbidity in preterm infants. Eur J Pediatr. 173(7):905-912.

9. Greenough A, Limb E, Marston L, Marlow N, Calvert S, Peacock J. 2005. Risk factors for respiratory morbidity in infancy after very premature birth. Archives of Disease in Childhood Fetal and Neonatal Edition. 90(4):F320-F323.

10. Keller RL, Feng R, DeMauro SB, Ferkol T, Hardie W, Rogers EE, Stevens TP, Voynow JA, Bellamy SL, Shaw PA et al. 2017. Bronchopulmonary dysplasia and perinatal characteristics predict 1-year respiratory outcomes in newborns born at extremely low gestational age: A prospective cohort study. The Journal of Pediatrics. 187:89-97.e83.

11. MacBean V, Drysdale S, Yarzi M, Peacock J, Rafferty G, Greenough A. 2018. Respiratory viral infections in infancy and school age respiratory outcomes and healthcare costs. Pediatr Pulmonol. 53(3):342-348.

12. Drysdale SB, Wilson T, Alcazar M, Broughton S, Zuckerman M, Smith M, Rafferty GF, Johnston SL, Greenough A. 2011. Lung function prior to viral lower respiratory tract infections in prematurely born infants. Thorax. 66(6):468-473.

13. Yüksel B, Greenough A. 1995. Functional residual capacity to thoracic gas volume (FRC:TGV) ratio in healthy neonates. Respir Med. 89(6):429-433.

14. Milner AD, Marsh MJ, Ingram DM, Fox GF, Susiva C. 1999. Effects of smoking in pregnancy on neonatal lung function. Arch Dis Child Fetal Neonatal Ed. 80(1):F8-14.

15. MacBean V, Jolley CJ, Sutton TG, Greeenough A, Moxham J, Rafferty GF. 2016. Parasternal intercostal electromyography: A novel tool to assess respiratory load in children. Pediatr Res. 80(3):407-414.

16. MacBean V, Wheatley L, Lunt AC, Rafferty GF. 2017. Respiratory load perception in overweight and asthmatic children. Resp Physiol Neurobiol 239: 81-86.

17. Department of Health. 2015. Reference costs 2014-15.

18. Joint Formulary Committee. 2016. British National Formulary (online). London: BMJ Group and Pharmaceutical Press.

19. Charrad M, Ghazzali N, Boiteau V, Niknafs A. 2014. An r package for determining the relevant number of clusters in a data set. 61:36.

20. Gardner S, le Roux N. 2005. Extensions of biplot methodology to discriminant analysis. J Classif. 22:59-86.

21. Strobl C, Malley J, Tutz G. 2009. An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. Psychol Methods. 14:323-348.

22. Beamer P, Lothrop N, Lu Z, Ascher R, Ernst K, Stern D, Billheimer D, Wright A, Martinez F. 2016. Spatial clusters of child lower respiratory illnesses associated with community-level risk factors. Pediatr Pulmonol. 51:633-642.

23. Dumas O, Mansbach JM, Jartti T, Hasegawa K, Sullivan AF, Piedra PA, Camargo CA. 2016. A clustering approach to identify severe bronchiolitis profiles in children. Thorax. 71(8):712-8

24. Rossi GA, Colin AA. 2015. Infantile respiratory syncytial virus and human rhinovirus infections: Respective role in inception and persistence of wheezing. Eur Respir J. 45(3):774-789.

25. Çalışkan M, Bochkov YA, Kreiner-Møller E, Bønnelykke K, Stein MM, Du G, Bisgaard H, Jackson DJ, Gern JE, Lemanske RF, Jr. et al. 2013. Rhinovirus wheezing illness and genetic risk of childhood-onset asthma. N Engl J Med. 368(15):1398-1407.

26. Aylward GP, Hatcher RP, Stripp B, Gustafson NF, Leavitt LA. 1985. Who goes and who stays: Subject loss in a multicenter, longitudinal follow-up study. J Dev Behav Pediatr. 6(1):3-8.

27. Williams PL, Van Dyke R, Eagle M, Smith D, Vincent C, Ciupak G, Oleske J, Seage GR, 3rd, Team PC. 2008. Association of site-specific and participant-specific factors with retention of children in a long-term pediatric HIV cohort study. Am J Epidemiol. 167(11):1375-1386.

LEGENDS

Table legends

Table 1: Baseline characteristics

Table 2 Distribution of viral infection frequencies in the first year across clusters. Data are presented as n (%).

Table 3 Healthcare utilisation costs across clusters from one year of age to point of school age follow up. Data are presented as mean (95% confidence intervals).

Figure 1: Decision metric, derived using conditional inference tree analysis, to allow classification of children into clusters based on a minimal subset of variables.

Supplementary file legends

Table 1: Characteristics of participants who did and did not participate in the school age follow up study.

Figure 1: Cluster dendrogram showing a three-cluster solution. The heights at which observations, or groups of observations, fuse indicate the degree of similarity. The dashed line denotes the cut point used to derive the clustering solution.

Figure 2: Discriminant projection plot of clustering solution. Shaded areas are the 95% confidence regions for each cluster assuming elliptical clusters

Table 1: Baseline characteristics

Data are presented as n or median (IQR)

	Cluster One	Cluster Two	Cluster Three	P value across clusters
n	140	20	8	
Sex (M : F)	80 : 60	12 : 8	4 : 4	0.89
Gestational age (weeks)	34^{+2} (32^{+6} to 35^{+2})	29^{+1} (28^{+2} to 31^{+2})	24^{+5} (23^{+6} to 26^{+3})	<0.001
Birth weight (g)	2042 (1708 to 2372)	1313 (1060 to 1508)	670 (628 to 732)	<0.001
Duration of invasive mechanical ventilation (days)	0 (0 to 1)	13 (8 to 17)	86 (73 to 101)	<0.001
Duration of supplemental oxygen (days)	0 (0 to 1)	23 (10 to 52)	119 (99 to 257)	<0.001
Functional residual capacity (z scores)	-0.67 (-2.06 to 0.73)	-2.13 (-2.71 to -0.04)	-5.12 (-6.13 to -4.64)	<0.001
Respiratory system resistance (z scores)	0.33 (-0.15 to 1.11)	1.18 (0.57 to 2.70)	2.41 (0.94 to 4.06)	0.009
Respiratory system compliance (z scores)	0.57 (-0.05 to 1.14)	-0.34 (-0.59 to 0.59)	-0.64 (-1.09 to -0.20)	<0.001

Table 2 Distribution of viral infection frequencies in the first year across clusters. Data are presented as n (%).

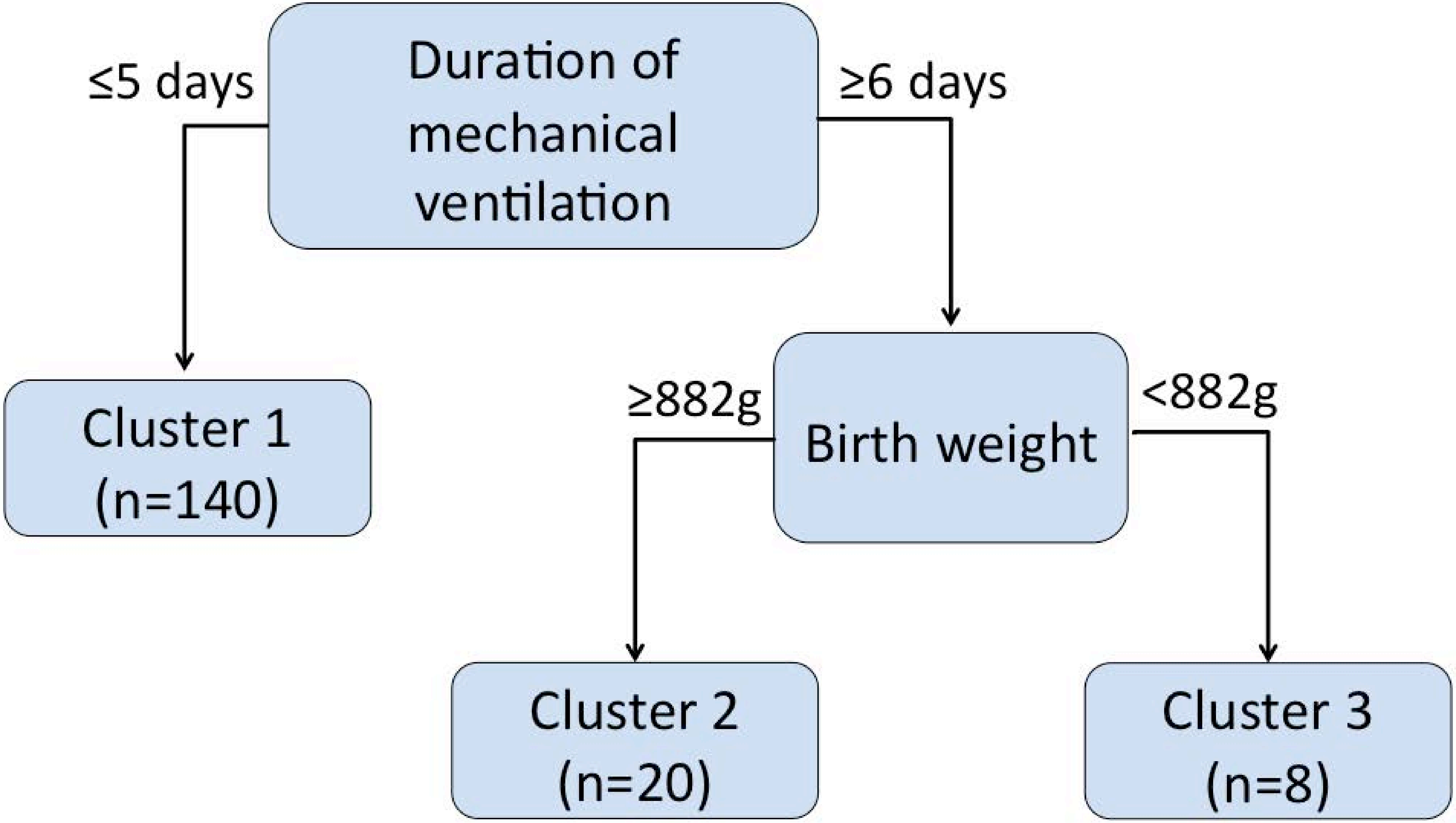
	Cluster one	Cluster two	Cluster three
n	128	17	6
LRTI in first year	62 (48%)	11 (65%)	6
No LRTI (n (%))	66 (51.6)	6 (35.3)	0 (0)
RSV LRTI (n (%))	21 (16.4)	7 (41.2)	1 (16.7)
RV LRTI (n (%))	15 (11.7)	1 (5.9)	4 (66.7)
Other LRTI (n (%))	26 (20.3)	3 (17.6)	1 (16.7)

Table 3 Healthcare utilisation costs across clusters from one year of age to point of school age follow up.

Data are presented as mean (95% confidence intervals).

	Cluster One	Cluster Two	Cluster Three	p value across clusters
n	45	6	5	
Overall respiratory healthcare costs (£/year)	110.54 (29.17 to 191.92)	100.74 (-65.46 to 266.95)	1575.05 (2396.96 to 5547.06)	0.008
Overall non-respiratory healthcare costs (£/year)	215.81 (116.78 to 314.84)	950.68 (70.07 to 1831.28)	1887.15 (1326.83 to 5101.14)	0.003
Respiratory related GP attendances (number/year)	0.89 (0.49 to 1.29)	1.17 (-0.38 to 2.71)	3.20 (0.24 to 6.16)	0.014
Non-respiratory related GP attendances (number/year)	1.12 (0.89 to 1.36)	3.69 (-1.24 to 8.61)	2.02 (0.70 to 3.34)	0.089
Respiratory medication costs (£/year)	6.44 (2.11 to 10.76)	11.25 (-6.76 to 29.25)	31.21 (-9.78 to 72.20)	0.029

Non-respiratory medication costs (£/year)	22.44 (1.35 to 43.52)	20.74 (-15.24 to 56.71)	1068.87 (-1859.95 to 3997.69)	0.289
Respiratory related hospital costs (£/year)	67.56 (2.96 to 132.16)	36.81 (-57.81 to 131.42)	1405.58 (-2419.43 to 5230.69)	0.030
Non-respiratory related hospital costs (£/year)	144.01 (53.27 to 234.75)	767.76 (-60.86 to 1596.39)	729.33 (-9.53 to 1468.20)	0.002


Duration of mechanical ventilation
≤5 days
≥6 days
Cluster 1 (n=140)
Birth weight
≥882g
<882g
Cluster 2 (n=20)
Cluster 3 (n=8)