University of Cape Town

Latent Class Modelling of Respiratory Outcomes in a South African Birth Cohort

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy in the

Department of Statistical Sciences

Author:
Carlyle McCready[1,2]

Supervisor:
Prof. Francesca Little[1]
Co-Supervisor:
Prof. Heather J. Zar[2]
Associate Prof. Freedom Gumedze[1]

1. University of Cape Town, South Africa, Department of Statistical Sciences
2. University of Cape Town, South Africa, Department of Paediatrics and Child Health, and SA-MRC Unit on Child & Adolescent Health

June 19, 2025

University of Cape Town

The copyright of this thesis vests in the author. No quotation from it or information derived from it is to be published without full acknowledgement of the source. The thesis is to be used for private study or non-commercial research purposes only.

Published by the University of Cape Town (UCT) in terms of the non-exclusive license granted to UCT by the author.

University of Cape Town

Declaration

I, Carlyle McCready, hereby declare that the work on which this thesis is based is my original work (except where acknowledgements indicate otherwise) and that neither the whole work nor any part of it has been, is being, or is to be submitted for another degree in this or any other university. I authorise the University to reproduce for the purpose of research either the whole or any portion of the contents in any manner whatsoever.

Signed by candidate

June 19, 2025

University of Cape Town

Abstract

Faculty of Science
Department of Statistical Sciences

Doctor of Philosophy

by Carlyle McCready
Supervised by Prof. Francesca Little, Assoc Prof. Freedom Gumedze, and Prof. Heather J. Zar

Latent Class Modelling of Respiratory Outcomes in a South African Birth Cohort

Background & Purpose: Early life is a key period determining long-term health, particularly in low-income and middle-income countries (LMICs) where negative exposures are common. This thesis focuses on two key respiratory outcomes: childhood wheezing and lung function. It examines (1) the underlying heterogeneity of wheezing and lung function measurements through the derivation of latent phenotype classes and (2) the impact of early-life lower respiratory tract infections (LRTIs) and other predictors on wheezing and lung function.

Methods: Data from the Drakenstein Child Health Study (DCHS) was used to critically evaluate various statistical approaches for identifying latent structures and modelling longitudinal trajectories, tracking 1143 children until school age. To identify wheezing phenotypes, partition around medoids (PAM) clustering tailored for longitudinal data, latent class analysis (LCA) and LCA with random effects were applied and evaluated. Lung function measurements were adjusted for height (as a proxy for body size) using a multiplicative model approach. Prior to employing latent class mixed effects models (LCMM) to identify latent classes in longitudinal lung function data, generalised additive models for location, scale, and shape (GAMLSS) and interrupted time series (ITS) approaches were used to model the impact of LRTIs and other risk factors on longitudinal trajectories. Well-established distributions were chosen within the GAMLSS framework to minimise overfitting and ensure robustness.

Findings: PAM successfully identified four clinically distinct wheezing phenotypes: never, early transient, late-onset, and recurrent, outperforming LCA, which faced challenges with class separation. Random-effect LCA improved the modelling of recurrent wheezing trends but faced limitations with rare or sparse data patterns. While investigating phenotype-specific risk factors, cross-validation proved valuable to ensure the robustness and stability of variable selection and

model estimates by evaluating model output across different data subsets. Turning to lung function development, by integrating a multiplicative model approach, the GAMLSS framework, and the ITS methodology, this thesis provided a comprehensive understanding of growth patterns. It captured both population-level trends and individual variability while accounting for step changes over time. Using LMS z-scores, LCMM effectively identified underlying features in longitudinal data, with two latent classes capturing the heterogeneity, a normal group and a smaller group of children who deviated from it. The multivariate LCMM approach captured shared latent processes influencing multiple dimensions of lung function, improving the interpretability of class assignments. Class separation analysis highlighted the need for careful selection of the number of latent classes, balancing interpretability and the ability to capture subtle data structures.

Conclusions & implications: This thesis reveals how various factors shape respiratory health and the diverse pathways children follow in LMICs. The distinct wheezing patterns in LMICs (compared to a UK cohort) highlight the need for region-specific health policies and call for further study of long-term effects in diverse settings.

Keywords: Respiratory syncytial virus (RSV), Drakenstein Child Health Study (DCHS), partition around medoids (PAM), latent class analysis (LCA), latent class mixed effect models (LCMM), generalised additive models for location, scale and shape (GAMLSS), interrupted time series (ITS), lower respiratory tract infections (LRTI).

List of publications

Chapters 3 and 4 are based on or contain published work by McCready et al. (2023) and McCready et al. (2024). Specifically, Chapter 3 includes findings related to identifying risk factors for each wheezing phenotype, McCready et al. (2023). Chapter 4 explores the effects of age and exposure to various risk factors on the development of lung function outcomes, McCready et al. (2024). The content has been adapted and expanded to fit the scope of this thesis.

[1] Carlyle McCready, Heather Zar, Shaakira Chaya, Carvern Jacobs, Lesley Workman, Zoltan Hantos, and Hall, Graham L. and Sly, Peter D. and Nicol, Mark P., Dan J. Stein, Anhar Ullah, Adnan Custovic, Francesca Little, and Diane M. Gray, Determinants of lung function development from birth through school age in a South African birth cohort. *The Lancet Child & Adolescent Health*, 2024.

[2] Carlyle McCready, Sadia Haider, Francesca Little, Mark P Nicol, Lesley Workman, Diane M Gray, Raquel Granell, Dan J Stein, Adnan Custovic, and Heather J Zar. Early childhood wheezing phenotypes and determinants in a South African birth cohort: longitudinal analysis of the Drakenstein Child Health Study. *The Lancet Child & Adolescent Health*, 7(2):127135, 2023.

Conference presentations

[1] McCready C, Haider Sadia, Little F, Workman L, Custovic A, Zar HJ. A critical review and comparison of statistical clustering techniques to derive wheezing phenotypes in a South African birth cohort study. International Biometric Conference. Latvia. 10-15 July 2022.

[2] McCready C, Haider S, Little F, Workman L, Gray DM, Nicol M, Custovic A, Zar HJ. Wheezing phenotypes and their early-life determinants in a South African birth cohort study. South African Statistical Association (SASA), 2021.

[3] McCready C, Haider S, Little F, Workman L, Gray DM, Nicol M, Custovic A, Zar HJ. Wheezing phenotypes and early-life determinants in a South African birth cohort study. Department of Paediatrics & Child Health Annual Research Day. 04 November 2020.

Acknowledgments and contributions

With regards to the paper *Early childhood wheezing phenotypes and determinants in a South African birth cohort: longitudinal analysis of the Drakenstein Child Health Study*, Carlyle McCready contributed to analyses and writing of the original draft, Sadia Haider contributed to the supervision and analyses. Prof. Francesca Little contributed to the methodology, supervision, and oversaw data analyses. Prof. Mark P Nicol led the microbiology aspects. Lesley Workman did data curation. Diane M Gray contributed to lung function methodology, supervision, and analyses. Raquel Granell provided access and contributed to the analysis of ALSPAC data. Prof. Dan J Stein devised and oversaw psychosocial measures. Prof. Adnan Custovic contributed to the conceptualisation, funding acquisition, methodology, and supervision. Prof. Heather J Zar contributed to the conceptualisation, funding acquisition, methodology, and supervision.

With regards to the paper *Determinants of lung function development from birth through school age in a South African birth cohort*, Carlyle McCready contributed to analyses and writing of the original draft, Prof. Francesca Little contributed to methodology, supervision, and oversaw data analyses. Shaakira Chaya contributed to data collection and analysis. Carvern Jacobs contributed to data collection. Lesley Workman did the data curation and formal analysis. Carlyle McCready, Lesley Workman, Carvern Jacobs, Shaakira Chaya, and Diane M Gray had access to and verified the existence of the raw data. Prof. Zoltan Hantos contributed to the methodology, data acquisition and analysis, and developed oscillometry aspects of the study. Prof. Graham L Hall contributed to the conception and methodology. Prof. Peter D Sly contributed to the conception of the study. Prof. Mark P Nicol led the microbiology aspects. Prof. Dan J Stein devised and oversaw psychosocial measures. Anhar Ullah contributed to the analyses. Prof. Adnan Custovic contributed to the conceptualisation, funding acquisition, methodology, and supervision. Prof. Heather J Zar contributed to the conceptualisation, funding acquisition, methodology, and supervision. Diane M Gray contributed to lung function conceptualisation, funding acquisition, methodology, supervision, and analysis.

Thank you to the children and families participating in the Drakenstein Child Health Study (DCHS). Appreciation is extended to the study staff, as well as the clinical and administrative staff of the Western Cape Government Department of Health and Wellness, for their support of the study.

Contents

1 Introduction **3**
- 1.1 Aim and motivation . . . 3
- 1.2 The Drakenstein Child Health Study . . . 6
  - 1.2.1 Setting . . . 6
  - 1.2.2 Study design and participants . . . 6
  - 1.2.3 Early-life risk factors, and cohort flow diagram . . . 7
  - 1.2.4 Ethic approval . . . 10
  - 1.2.5 Data lock . . . 10
- 1.3 Outline . . . 10
- 1.4 Key contributions . . . 11

2 Literature review **13**
- 2.1 Aim and structure . . . 13
- 2.2 Childhood wheeze . . . 13
- 2.3 Lung function development in early life . . . 17

3 The identification of wheezing phenotypes in early childhood using binary indicators **23**
- 3.1 Introduction . . . 23
  - 3.1.1 Chapter aims . . . 23
- 3.2 Data sources and definition of variables . . . 24
- 3.3 Longitudinal visualisations of wheezing phenotypes . . . 25
- 3.4 Distance-based clustering . . . 26
  - 3.4.1 Distance measures . . . 26
  - 3.4.2 Partitioning methods . . . 29
- 3.5 Extending the PAM clustering approach . . . 33
  - 3.5.1 Multi-dimensional indicators . . . 33
  - 3.5.2 Procedure . . . 34
- 3.6 Model based clustering . . . 36
- 3.7 Results: Wheezing phenotypes in early childhood . . . 40

3.7.1 Partitioning around mediods (PAM) and multi-dimensional indicators 40
3.7.2 Sensitivity analysis of the 4-class phenotype solution under PAM 45
3.7.3 DCHS versus Avon Longitudinal Study of Parents and Children (ALSPAC) . 47
3.7.4 Latent class analysis (LCA) 50
3.8 Phenotype stability 56
3.9 Comparison of input variables for PAM and LCA 59
3.10 Risk factors for wheezing phenotypes identified using the PAM approach 61
3.10.1 Introduction 61
3.10.2 Identifying potential risk factors using a DAG 61
3.10.3 Descriptions of risk factors and their encoding 62
3.10.4 Statistical analysis methods 64
3.10.5 Clinical results 66
3.10.6 Cross validation framework for variable selection and stability 75
3.11 Discusion 84

4 Distinctive lung function trajectories from birth to 5-years in a South African birth cohort study 89
4.1 Introduction 89
4.1.1 Chapter aims 89
4.1.2 Details of DCHS lung function testing 91
4.2 Methods 92
4.2.1 GAMLSS framework 93
4.2.2 The LMS method 96
4.2.3 The Box-Cox T Method 97
4.2.4 Model selection 98
4.2.5 Interrupted time series (ITS) regression 99
4.2.6 Impact of height on lung function 100
4.2.7 Latent class mixed models (LCMM) 102
4.3 Results 105
4.3.1 Understanding developmental trajectories of childhood lung function 105
4.3.2 Phases in lung function trajectories from birth 109
4.3.3 The impact of height on lung function development 111
4.3.4 Modelling the height-adjusted lung function outcomes 120
4.3.5 Determinants of lung function development 128
4.3.6 Clinical results 129
4.3.7 Latent profiles of longitudinal lung function outcomes 136
4.3.8 A feature-based approach to cluster longitudinal lung function 149
4.3.9 Early life risk factors and latent classes 157
4.4 Discussion 167

5 Conclusion and unifying wheezing and lung function phenotypes insights through correspondence analysis **173**
5.1 Introduction . . . 173
5.1.1 Concluding summary . . . 173
5.1.2 Bringing the different chapters together . . . 175
5.1.3 Multiple correspondence analysis (MCA) . . . 175
5.1.4 Discussion . . . 176

A **181**
A.1 Chapter 3 . . . 181
A.2 Chapter 4 . . . 183
A.2.1 Multivariate LCMM outcomes for oscillometry outcomes . . . 191
A.2.2 Multivariate LCMM outcomes for MBW and tidal breathing measures . . . 194
A.2.3 Univariate LCMM outcomes for oscillometry outcomes . . . 198
A.2.4 Univariate LCMM outcomes for MBW and tidal breathing outcomes . . . 201

List of Abbreviations or Acronyms

AIC	Akaike information criterion
ASSIST	Alcohol, smoking and substance involvement screening test
ASW	Average silhouette width
AOS	Airwave oscillometry
AV	Adenovirus
BCCG	Box-Cox Cole and Green
BCPE	Box-Cox Power exponential
BCT	Box-Cox t
BIC	Bayesian information criterion
BMI	Body mass index
bs	Basis spline
CIs	Confidence intervals
DCHS	Drakenstein Child Health Study
DTW	Dynamic time warping
ELISA	Enzyme-linked immunosorbent assay
EPDS	Edinburgh postnatal depression scale
ETW	Early transient wheeze
FEF	Forced expiratory flow
FEV1	Forced expiratory volume
FRC	Functional residual capacity
FVC	Forced vital capacity
GAM	Generalised additive model
GAMLSS	Generalised Additive Models for Location Scale and Shape
GD	Global deviance
GINA	Global Initiative for Asthma
GLME	Generalised linear mixed effects
GMM	Growth mixture model
Hib	Haemophilus influenzae type b
HICs	High-income countries
HIV	Human immunodeficiency virus

IPV	Intimate partner violence
ISAAC	International Study of Asthma and Allergies in Childhood
ITS	Interrupted time series
LCA	Latent class analysis
LCI	Lung clearance index
LCMM	Latent class mixed modelling
LME	Linear mixed effects
LMICs	Low-income and middle-income countries
LMS	Lambda, mu, and sigma
LOW	Late onset wheeze
LRTI	Lower respiratory tract infection
MADAM	Mean and dispersion additive model
MCA	Multiple correspondence analysis
ML	Maximum likelihood
NO	Normal
NWZ	Never wheeze
PAM	Partition around medoids
PCR	Polymerase chain reaction
Ree	Respiratory resistance at the end of expiration
Rrs	Respiratory system resistance
RS	Rigby and Stasinopoulos
RSV	Respiratory syncytial virus
RV	Rhinovirus
RW	Recurrent wheeze
SARP	Severe Asthma Research Program
SCA	Simple correspondence analysis
SES	Socio-economic status
SOM	Self-organising maps
SRQ20	Self-reported questionnaire 20-item
TBFVL	Tidal breathing flow volume loops
$\mathbf{t_{PTEF}/t_E}$	Ratio time of peak total expiratory flow to time of expiration
TF	T-family
UCT	University of Cape Town
UK	United Kingdom
WHO	World Health Organisation
Xee	Respiratory reactance at the end of expiration
Xrs	Respiratory system reactance
Zrs	Respiratory impedance

List of Tables

1.1 Respiratory outcomes of interest from the DCHS 5
1.2 Definitions of child, maternal, and socio-economic characteristics from the DCHS . . 9

3.1 Derivation of wheezing indicators 34
3.2 Four classes of latent variable methods by Bartholomew and Knott (1999) 36
3.3 Average silhouette width for various phenotype sizes identified by the PAM model in Figure 3.5. 41
3.4 Distribution of the derived indicators stratified by phenotype (PAM 4-Class model). 44
3.5 Wheezing classes identified by the PAM model - Average silhouette width for indicator subsets 46
3.6 BIC values for LCA and LCA RE models. 50
3.7 A 2 by 2 cross-table of class assignments between LCA and random LCA (Common) models. 52
3.8 Top fifteen wheeze patterns in the DCHS 53
3.9 Wheezing phenotypes from the PAM model: Average silhouette width and standard deviation for subsets across iterations 57
3.10 Wheezing phenotypes from the LCA model: Average BIC and standard deviation for subsets across iterations 58
3.11 Distribution of lower respiratory illness in DCHS children from birth to age 5 years. 67
3.12 Prevalence of RSV, RV, AV, influenza, or parainfluenza-LRTI, by wheezing phenotype. 67
3.13 Unadjusted multinomial logistic regression of the association of early-life factors with wheezing phenotypes in the DCHS (reference class: Never wheezing) 69
3.14 Adjusted multinomial logistic regression of factors associated with wheezing phenotypes in DCHS (reference class: Never wheezing) 70
3.15 Association of lung function (oscillometry) at 5 years and wheeze phenotypes in DCHS. 71

4.1 Summary of lung function measurements by visit. 106
4.2 Effect of log(height) on lung function outcomes: μ Coefficients of log(Height) from Equation 4.16 111

4.3 Model formulation used to illustrate the similarity of coefficients while modelling the unadjusted or the height-adjusted FRC outcomes . . . 118
4.4 Comparison of model coefficients between unadjusted FRC (Model 1) and height-adjusted FRC outcomes (Model 2) from Table 4.3 . . . 119
4.5 GAMLSS model summary for height-adjusted FRC (linear model) . . . 122
4.6 Trajectories of multiple breath washout, tidal breathing and oscillometry outcomes from birth . . . 123
4.7 Model fit statistics for lung function outcomes . . . 124
4.8 Fitted models for each distribution assumed in GAMLSS . . . 125
4.9 Association of early life risk factors with the development of multiple breath washout and tidal breathing parameters from birth to age 5 years; multivariate model. . . . 130
4.10 Association of early life risk factors with the development of oscillometry parameters from birth to 5 years; multivariate model . . . 131
4.11 Association of maternal HIV disease severity with LCA, tidal volume, and Xee from birth to 5 years. . . . 132
4.12 Comparison of LRTI incidence between HIV exposed uninfected and HIV unexposed groups . . . 132
4.13 Lung function outcomes by LRTI history: never versus ever diagnosed . . . 133
4.14 Summary table: Univariate airwave oscillometry spline models depicted in Figure 4.16 142
4.15 Summary table: Univariate MBW and tidal breathing spline models depicted in Figure 4.18 . . . 144
4.16 Summary table: Multivariate airwave oscillometry, MBW, and tidal breathing spline models . . . 148
4.17 Average silhouette width (ASW) values obtained from step 2 using the PAM algorithm for linear, quadratic, and spline models from step 1 in the two-stage feature-based approach for compliance. . . . 153
4.18 Summary table, for compliance, of the class sizes for scenario 1, 3, and 5 in Table 4.17. 153
4.19 Average silhouette width (ASW) values obtained from step 2 using the PAM algorithm for linear, quadratic, and spline models from step 1 in the two-stage feature-based approach for Xee. . . . 154
4.20 Summary table, for Xee, of the class sizes for scenario 1, 3, and 5 in Table 4.19. . . . 154
4.21 Summary table: LCMM results for z-score resistance, Ree, Xee, and compliance models 158
4.22 Model estimates for resistance, compliance, Ree, and Xee from LCMM. . . . 159
4.23 Summary table: LCMM results for z-score FRC, LCI, RR, tidal volume, and ratio models . . . 160
4.24 Model estimates for FRC, LCI, respiratory rate, tidal volume, and t_{PTEF}/t_E from LCMM. . . . 161

A.1 Comparison of included versus excluded children in the DCHS . . . 181
A.2 Bayesian Information Criterion (BIC) values for different phenotype sizes which were used by the LCA model to identify the wheezing phenotypes in Figure 3.9. . . . 182

A.3 Association of specific LRTI with the development of multiple breath washout and tidal breathing parameters from birth to 5 years 190
A.4 Association of LRTIs with the development of oscillometry parameters from birth to age 5 years. 190
A.5 Summary table: Multivariate airwave oscillometry linear and quadratic models depicted in Figure A.5 and A.6 193
A.6 Summary table: Multivariate MBW and tidal breathing linear, quadratic, and spline models depicted in Figure A.7, A.8, and A.9 197
A.7 Summary table: Univariate airwave oscillometry linear models depicted in Figure A.10200
A.8 Summary table: Univariate airwave oscillometry quadratic models depicted in Figure A.11 200
A.9 Summary table: Univariate MBW and tidal breathing linear models depicted in Figure A.12 203
A.10 Summary table: Univariate MBW and tidal breathing quadratic models depicted in Figure A.13 203

List of Figures

1.1 Cohort flow diagram. 8

3.1 Scheduled Visits 24
3.2 Point prevalence of current wheeze in children in the DCHS by age. 25
3.3 The format of the primary input data matrix (left), multidimensional indicators (centre), and the distance matrix of pairs (right) under the PAM approach. 35
3.4 Colour coding used in the heatmaps. 41
3.5 Percentage of children with reported wheeze in the first 5 years of life within each wheeze phenotype (left). Intra-phenotype individual wheeze patterns (right), stratified by phenotype; the darker colours of the bars or lines represent the presence of wheeze and lighter colours represent the absence of wheeze. 42
3.6 Point prevalence of current wheeze in children in the ALSPAK by age. See McCready et al. (2023). 47
3.7 ALSPAC cohort - Percentage of children with reported wheeze within each wheeze phenotype (left). Intra-phenotype individual wheeze patterns (right), stratified by phenotype; the darker colours of the bars or lines represent the presence of wheeze and lighter colours represent the absence of wheeze. See McCready et al. (2023). . . 48
3.8 DCHS cohort 5-class solution - Percentage of children with reported wheeze within each wheeze phenotype (left). Intra-phenotype individual wheeze patterns (right), stratified by phenotype; the darker colours of the bars or lines represent the presence of wheeze and lighter colours represent the absence of wheeze. 49
3.9 Identified by the LCA model. Percentage of children with reported wheeze in the first 5 years of life within each wheeze phenotype (left). Intra-phenotype individual wheeze patterns (right), stratied by phenotype; the darker colours of the bars or lines represent the presence of wheeze and lighter colours represent the absence of wheeze. 51
3.10 Percentage of children with reported wheeze in the first 5 years of life within each wheeze phenotype (left). Intra-phenotype individual wheeze patterns (right), stratied by phenotype; the darker colours of the bars or lines represent the presence of wheeze and lighter colours represent the absence of wheeze. 52

3.11 A random LCA example with low prevalence: Percentage of children with reported wheeze in the first 5 years of life within each wheeze phenotype (left). Intra-phenotype individual wheeze patterns (right), stratified by phenotype; the darker colours of the bars or lines represent the presence of wheeze and lighter colours represent the absence of wheeze. . . . 54
3.12 A random LCA example with higher prevalence: Percentage of children with reported wheeze in the first 5 years of life within each wheeze phenotype (left). Intra-phenotype individual wheeze patterns (right), stratified by phenotype; the darker colours of the bars or lines represent the presence of wheeze and lighter colours represent the absence of wheeze. . . . 55
3.13 Profiles of wheeze phenotypes over time - Analysis of class stability with respect to changes in different sample sizes[1] of data . . . 57
3.14 Profiles of wheeze phenotypes over time - Analysis of class stability with respect to changes in different sample sizes. . . . 58
3.15 Percentage of children with reported wheeze in the first 5 years of life within each wheeze phenotype (left). Intra-phenotype individual wheeze patterns (right). Application of PAM to binary wheezing outcomes. . . . 59
3.16 Directed acyclic graph (DAG) for the evaluation of risk factors in the multinomial logistic regression model . . . 62
3.17 Airway resistance at ages 6 weeks and 5 years in four wheeze clusters in DCHS (mean z-scores, 95% CI). NWZ = Never wheeze; ETW = Early transient wheeze; LOW = Late onset wheeze; RW = Recurrent wheeze . . . 72
3.18 Probability of wheezing for a phenotype versus the continuous Xee outcome (left) and Ree (right). . . . 73
3.19 Analysis of residuals plots assessing linearity of data (left), homogeneity of residuals variance (centre), and normality of residuals (right) . . . 74
3.20 A cross-validation (CV) framework for variable selection and stability . . . 76
3.21 Cross-validation for variable selection . . . 78
3.22 Stability analysis sex using cross-validation. Odds ratios (left) and p-values (right). . 79
3.22 Stability analysis sex using cross-validation. Odds ratios (left) and p-values (right). . 80
3.22 Stability analysis sex using cross-validation. Odds ratios (left) and p-values (right). . 81
3.23 Density plots of odds ratio estimates by sample size. . . . 82
3.23 Density plots of odds ratio estimates by sample size. . . . 83

4.1 Relationship between age, height, and lung function outcomes: FRC (left) and LCI (right) as exemplars. . . . 90
4.2 Annual lung function assessments. . . . 91
4.3 Standardisation methods for lung function outcomes with FRC as an example. . . . 93
4.4 Illustration of the interrupted time series design . . . 99

4.5 Longitudinal plots of measured lung function outcomes (a) functional residual capacity, (b) lung clearance index, (c) resistance, (d) compliance, (e) ratio time of peak total expiratory flow to time of expiration (t_{PTEF}/t_E), and (f) tidal volume versus age. The dotted line between 2 years and 3 years highlights the step change between testing methods. 107

4.5 Longitudinal plots of measured lung function outcomes (g) respiratory rate, (h) respiratory resistance at the end of expiration (Ree), and (i) respiratory reactance at the end of expiration (Xee) versus age. The dotted line between 2 years and 3 years highlights the step change between testing methods. 108

4.6 Quartile plots of measured lung function outcomes (a) functional residual capacity, (b) lung clearance index, (c) resistance, (d) compliance, (e) ratio time of peak total expiratory flow to time of expiration (t_{PTEF}/t_E), (f) tidal volume, (g) respiratory rate, (h) respiratory resistance at the end of expiration (Ree), and (i) respiratory reactance at the end of expiration (Xee) versus age. 110

4.7 FRC measurements against $Y' = FRC/Height^b$ (left) and LCI measurements against $Y' = LCI/Height^b$ (right) 112

4.8 Longitudinal plots of height-adjusted lung function outcomes (a) functional residual capacity, (b) resistance, (c) compliance, (d) tidal volume, (e) respiratory rate, (f) respiratory resistance at the end of expiration (Ree), and (g) respiratory reactance at the end of expiration (Xee) versus age. The dotted line between 2 years and 3 years highlight the step change between testing methods. 113

4.9 (a) FRC measurements against $Y' = FRC/Height^b$ and (b) LCI measurements against $Y' = LCI/Height^b$. The first and third columns represent unadjusted Y, while the second and fourth columns present height-adjusted Y′. 114

4.9 (c) Resistance measurements against $Y' = Resistance/Height^b$ and (d) Compliance measurements against $Y' = Compliance/Height^b$. The first and third columns represent unadjusted Y, while the second and fourth columns present height-adjusted Y′. 115

4.9 (e) Tidal volume measurements against $Y' = Tidal\ volume/Height^b$ and (f) Respiratory rate measurements against $Y' = Respiratory\ rate/Height^b$. The first and third columns represent unadjusted Y, while the second and fourth columns present height-adjusted Y′. 116

4.9 (g) Ree measurements against $Y' = Ree/Height^b$ and (h) Xee measurements against $Y' = Xee/Height^b$. The first and third columns represent unadjusted Y, while the second and fourth columns present height-adjusted Y′. 117

4.10 Observed data points alongside fitted linear, quadratic, and spline models showcasing each model's ability to represent the underlying pattern. For each measured lung function outcome (a) functional residual capacity, (b) lung clearance index, (c) Ratio time of peak total expiratory flow to time of expiration (t_{PTEF}/t_E), (d) tidal volume, (e) respiratory rate versus age. 126

4.11 Observed data points alongside fitted linear, quadratic, and spline models showcasing each model's ability to represent the underlying pattern. For each measured lung function outcome (a) resistance, (b) compliance, (c) Ree, and (d) Xee versus age. . . 127
4.12 DAG for evaluation of exposures. 128
4.13 Association of specified LRTIs with the development of lung function parameters from birth to 5 years. 134
4.14 Predicted height-adjusted resistance and compliance by exposure groups over time. 135
4.15 Analysis approach to univariate and multivariate longitudinal outcomes for airwave oscillometry 137
4.16 This figure illustrates the heterogeneous structures within longitudinal univariate airwave oscillometry outcomes using spline models. Points show individual scores, while classes were computed through the loess function in R for visualisation purposes. 140
4.17 Cross tabulations and Sankey diagrams illustrates how the heterogeneous structures developed within longitudinal univariate airwave oscillometry outcomes, using spline models, as we increase the number of classes from 1 to 4. 141
4.18 This figure illustrates the heterogeneous structures within longitudinal univariate MBW (left) and tidal breathing (right) outcomes using spline models. Points show individual scores, while classes were computed through the loess function in R for visualisation purposes. 143
4.19 This figure illustrates the heterogeneous structures within longitudinal multivariate airwave oscillometry outcomes using spline models. Points show individual scores, while classes were computed through the loess function in R for visualisation purposes. 146
4.20 This figure illustrates the heterogeneous structures within longitudinal multivariate MBW and tidal breathing outcomes using spline models. Points show individual scores, while classes were computed through the loess function in R for visualisation purposes. 147
4.21 Illustration of trajectory clustering in lung function outcomes using compliance as an example. (a) The LCMM analysis from Figure 4.17 shows minimal deviation from a single trajectory. (b) In contrast, the feature-based approach aims to produce at least two distinct, well-defined trajectories of sufficient size. 149
4.22 This figure illustrates the heterogeneous structures within compliance z-scores using the two-stage feature-based approach with spline models. Points show individual scores, while classes were computed through a loess function for visualisation purposes. 155
4.23 Cross tabulations and Sankey diagrams illustrates how the heterogeneous structures developed within the compliance outcome, using a feature-based approach, as we increase the number of classes from 1 to 4. 155
4.24 This figure illustrates the heterogeneous structures within Xee z-scores using the two-stage feature-based approach with spline models. Points show individual scores, while classes were computed through a loess function for visualisation purposes. . . . 156

4.25 Cross tabulations and Sankey diagrams illustrates how the heterogeneous structures developed within the Xee outcome, using a feature-based approach, as we increase the number of classes from 1 to 4. 156
4.26 This figure illustrates the heterogeneous structures within airwave oscillometry z-scores using LCMM with spline models. 164
4.27 This figure illustrates the heterogeneous structures within MBW and tidal breathing resistance z-scores using LCMM with spline models. 165

5.1 MCA biplots for (a) MBW, (b) airwave oscillometry, and (c) tidal breathing measures 177

A.1 Plot of average silhouette width in ALSPAC to determine optimal number of clusters using the PAM algorithm. 182
A.2 Plot of average silhouette width to determine optimal number of clusters using the PAM algorithm - application of PAM to binary wheezing outcomes 182
A.2 Density histograms with theoretical densities of normal, BCCG, BCT distributions overlaid for each measured lung function outcome (a) functional residual capacity, (b) lung clearance index, (c) resistance, (d) compliance, (e) Ratio time of peak total expiratory flow to time of expiration (t_{PTEF}/t_E), (f) tidal volume, (g) respiratory resistance at the end of expiration (Ree), and (g) respiratory reactance at the end of expiration (Xee). 187
A.3 Unstandardised lung function outcomes plotted against the height-adjusted outcomes (Y') for (a) resistance, (b) compliance, (c) tidal volume, (d) respiratory rate, (e) respiratory resistance at the end of expiration (Ree), and (f) respiratory reactance at the end of expiration (Xee) 188
A.4 Lung function outcomes by LRTI history: never versus ever diagnosed. A visual representation of the medians with interquartile ranges (IQR) shown in Table 4.13 . 189
A.5 This figure illustrates the heterogeneous structures within longitudinal multivariate airwave oscillometry outcomes using linear models. Points show individual scores, while classes were computed through the loess function in R for visualisation purposes. 191
A.6 This figure illustrates the heterogeneous structures within longitudinal multivariate airwave oscillometry outcomes using quadratic models. Points show individual scores, while classes were computed through the loess function in R for visualisation purposes. 192
A.7 This figure illustrates the heterogeneous structures within longitudinal multivariate MBW (left) and tidal breathing (right) outcomes using linear models. Points show individual scores, while classes were computed through the loess function in R for visualisation purposes. 194
A.8 This figure illustrates the heterogeneous structures within longitudinal multivariate MBW (left) and tidal breathing (right) outcomes using quadratic models. Points show individual scores, while classes were computed through the loess function in R for visualisation purposes. 195

A.9 This figure illustrates the heterogeneous structures within longitudinal multivariate MBW (left) and tidal breathing (right) outcomes using spline models. Points show individual scores, while classes were computed through the loess function in R for visualisation purposes. 196
A.10 This figure illustrates the heterogeneous structures within longitudinal univariate airwave oscillometry outcomes using linear models. Points show individual scores, while classes were computed through the loess function in R for visualisation purposes. 198
A.11 This figure illustrates the heterogeneous structures within longitudinal univariate airwave oscillometry outcomes using quadratic models. Points show individual scores, while classes were computed through the loess function in R for visualisation purposes. 199
A.12 This figure illustrates the heterogeneous structures within longitudinal univariate MBW (left) and tidal breathing (right) outcomes using linear models. Points show individual scores, while classes were computed through the loess function in R for visualisation purposes. 201
A.13 This figure illustrates the heterogeneous structures within longitudinal univariate MBW (left) and tidal breathing (right) outcomes using quadratic models. Points show individual scores, while classes were computed through the loess function in R for visualisation purposes. 202

Chapter 1

Introduction

1.1 Aim and motivation

The study of respiratory outcomes in early life is crucial in low- and middle-income countries (LMICs) as it enhances our comprehension of lung development, enables the identification of children at risk for respiratory diseases, and facilitates the development of innovative strategies for the prevention and treatment of respiratory problems. The identification of children at risk for respiratory diseases in LMICs is essential for early intervention and preventive measures, helping to alleviate the overall burden of these diseases. Of interest in this study is that LMICs may have unique environmental and socio-economic factors that impact lung development differently than in high-income countries (HICs).

The importance of studying respiratory outcomes is seen in both HICs and LMICs as respiratory diseases are a prevalent cause of illness and mortality in children worldwide, [1]. There has been a substantial reduction in under-5 mortality since the beginning of 1990, [2]. Although this is a promising statistic, the burden of respiratory diseases is still heavily felt in LMICs, [3]. The significance of these observations underscores the need for a comprehensive investigation into respiratory outcomes within specific populations, which is a focus of this study.

Clinically, this study addresses and identifies underlying longitudinal latent profiles of respiratory outcomes within a well-defined African birth cohort. Additionally, the study aims to investigate the risk factors associated with each latent profile to provide an understanding of respiratory health dynamics in this population. In this study, *respiratory outcomes* refer specifically to childhood wheeze and childhood lung function development. Secondly, from a statistical perspective, this study investigates different methodological approaches used to identify latent profiles through a critical review, application, and comparison of these methods. The suitability of distance-based and model-based clustering approaches is assessed. The statistical approaches were selected to include both an established gold-standard method commonly used in high-income settings and newer, flexible alternatives that have been applied in comparable longitudinal cohorts. An overarching goal is to

blend a selection of complex modelling approaches.

It is hypothesised that observed disease patterns (or similarly, developmental trajectories of a child) can be described through the existence of at least two unobserved latent variables (or classes). The latent variables would represent subgroups of children from a selected cohort who have similar disease symptoms, patterns, or severity over time. In a medical research setting, latent variables can be described as the "true state of a disease"[4]; which could (in a simple scenario) be dichotomously represented through a disease-negative class and a disease-positive class. Here, the disease-positive class could possibly be further subdivided into an ambulatory-disease-positive and a hospitalised-disease-positive class based on severity. More explicitly, the outcome of interest is a categorical latent variable (with an optimally selected number of levels) that uncovers heterogeneous patterns in the temporal variation and severity of diseases across individuals.

The idea that specific disease labels are needed when addressing patients on an individual level, in place of an umbrella term, has been recommended by medical professionals specifically in the Lancet Commission titled *Redefining asthma as an airways disease*[5], where wheezing underlies a spectrum of airway diseases that ranges from transient symptoms with a lower respiratory tract infection (LRTI) to recurrent episodes in asthma. In this thesis, it is believed that this idea has the potential to be extended to other respiratory outcomes.

The identification of longitudinal developmental trajectories is a core aim of this research. The respiratory outcomes of interest for this analysis are summarised in Table 1.1. The different outcome types (binary versus continuous), allow different statistical frameworks to be investigated which can be used to analyse clinical data collected through either parental self-report (wheeze) or through clinical hardware and systems (lung function).

Wheeze, a common symptom of asthma, is particularly concerning in children in LMICs due to the burden they place on resource-limited healthcare systems. The prevalence of wheezing in African and notably South African children has been increasing, [6, 7]. Childhood wheezing manifests itself as a continuous whistling noise that is heard during inspiration or expiration as air moves through restricted bronchial passages, [8]. Wheezing leads to an audible sound that can often be heard by both the patient and the physician. This sound is a clinical indicator of underlying airway obstruction.

Lung development during childhood predicts lung function achieved in adulthood and morbidity and mortality throughout life. In Table 1.1, the headings multiple breath washout (MBW), Tidal breathing, and Airwave oscillometry (AOS) refer to different lung function test measures. MBW testing is used to assess how efficiently gas mixes in the lungs during tidal breaths and is useful for detecting early changes or impairment in lung function. LCI and FRC are the main outcomes of MBW testing. MBW testing is a valuable tool for assessing lung function in pediatric patients. It is well-tolerated by young children and does not require forced breathing maneuvers, [9]. Tidal breathing, the process of inhalation and exhalation during restful breathing, is important to consider

when evaluating lung disease. This is because lung disease can alter both the typical pattern of breathing and the way breathing changes over a series of breaths. Airwave oscillometry is a technique that can be used to estimate airway resistance and reactance from the central or distal airways. This technique requires a low level of cooperation and is done during normal breathing without forced maneuvers. Airwave oscillometry has been successfully used to measure lung function in preschool children, who are commonly unable to perform spirometry, [10]. The intra-breath analysis provides novel measures which are believed to offer greater sensitivity in identifying airway diseases or obstruction, [11].

This study investigates and develops methods to cluster children into subgroups and then find the differences between these groups. Through clustering, researchers can study the association between disease patterns and demographic factors, the severity of a disease, the impact of early-life viral pathogens and other maternal and child risk factors. Identifying specific risk factors may allow for the possible development of improved diagnostic, preventive, and management strategies for children, particularly in LMICs.

While investigating longitudinal latent profiles, the goal is not to find a single model of best fit based solely on fit statistics but rather to explain the heterogeneous structures within the data by exploring different latent class structures.

Table 1.1: Respiratory outcomes of interest from the DCHS

Outcome description	**Outcome type**
Childhood Wheeze	
Wheeze prevalence	Binary
Lung function development	
Multiple breath washout	
Functional residual capacity (Litres)	Continuous
Lung clearance index (number of FRC turnovers)	Continuous
Tidal breathing	
Tidal volume (millilitres)	Continuous
Respiratory rate (breaths per minute)	Continuous
Time to peak tidal expiratory time over total expiratory time (%)	Continuous
Airwave oscillometry	
Resistance ($hPa.s.L^{-1}$)	Continuous
Compliance ($L.hPa^{-1}$)	Continuous
Ree ($hPa.s.L^{-1}$)	Continuous
Xee ($hPa.s.L^{-1}$)	Continuous

Ree = respiratory resistance at the end of expiration
Xee = respiratory reactance at the end of expiration

1.2 The Drakenstein Child Health Study

1.2.1 Setting

The Drakenstein Child Health Study (DCHS), is a birth cohort study that aims to investigate the epidemiology, risk factors, aetiology and long-term impact of early LRTI on child health and provides vital insight into child health development in LMICs, [12, 13]. The main study sites were located in Paarl, a city within the Drakenstein Local Municipality, in the Western Cape province of South Africa. Continuous surveillance was implemented at the Paarl hospital and two local clinics in the area: the Mbekweni and TC Newman clinics.

The community, which generally consists of low-income households, is stable, with low levels of immigration or emigration. The population size of the Drakenstein Municipality is approximately 294 000 individuals, with 68 000 households and a child population of around 72 000, [14]. As of 2020, the local population had access to 17 primary healthcare clinics (11 fixed and 6 mobile clinics), in addition to 1 regional hospital and 3 community daycare centres, [14]. Overall, the Drakenstein region has a well-established primary healthcare system that provides for the majority of the population.

Traditionally, most studies investigating early child health and development within a well-defined cohort have taken place in higher-income countries, [15, 16]. This creates a need for a study that allows us to capitalise on a unique and well-defined African birth cohort. The children enrolled in the DCHS cohort develop in a resource-limited setting which is representative of many South African communities.

1.2.2 Study design and participants

All births occurred at a single public hospital in Paarl, South Africa, where birth parameters were obtained by study staff. The study visits were scheduled to coincide with immunisation visits (diphtheria, tetanus, acellular pertussis, Hib, influenzae and inactivated polio vaccine at 6, 10, 14 weeks and 18 months, measles vaccine at 9 and 18 months and 13-valent PCV at 6 weeks, 14 weeks and 9 months); additional study visits were done 6 monthly from birth through 5 years. Mother-child pairs could also participate in intensive 2 weekly follow-ups during the first year, [12].

The study enrolled pregnant women from 5 March 2012 to 31 March 2015 during their second trimester of pregnancy at the two public health clinics. Inclusion criteria required pregnant women to be 18 years or older, 20-28 weeks gestation, and a resident in the area. These participants are predominantly of low socioeconomic status (SES) with a high prevalence of risk factors which are common in LMIC settings and which negatively impact the development of respiratory outcomes, including a high burden of infectious diseases and poor living conditions.

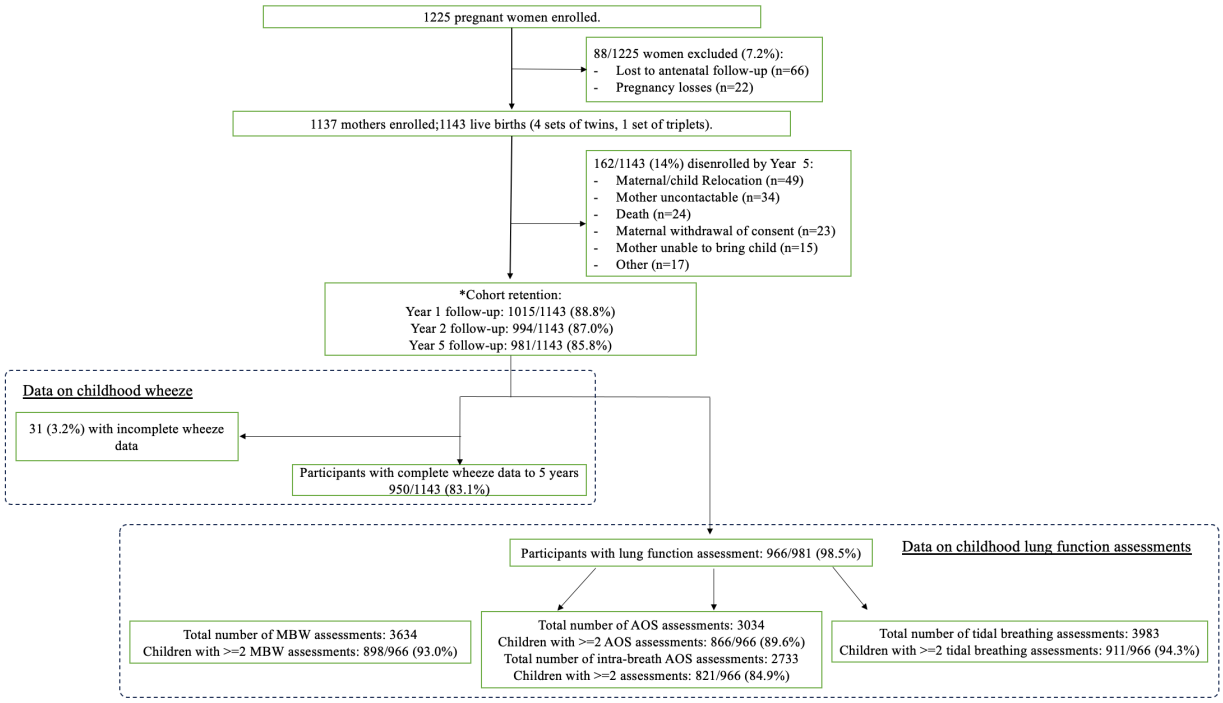
1.2.3 Early-life risk factors, and cohort flow diagram

Common risk factors include environmental exposures (air pollution or tobacco smoke), immunological factors (human immunodeficiency virus (HIV) exposure, atopy), maternal health (atopy, tobacco smoke, psychological stress) and social deprivation (overcrowding, food insecurity, poor maternal education). These antenatal and early life risk factors are associated with diminished lung growth and development through childhood. For example, LRTI in early childhood has been associated with impairment in lung function through the life course[17] and premature mortality[18], but most studies have been done in HICs, [17, 19]. Impaired lung function in childhood is associated with low lung function in adulthood, [20]. Known risk factors of childhood wheezing include early-life LRTIs[21], with respiratory syncytial virus (RSV), which is frequently associated with recurrent wheezing in high-income countries, being a predominant causal pathogen, [22]. Furthermore, high exposure to air pollution, tobacco smoke, psychosocial stressors, or poor living conditions might lead to different patterns of wheezing illness in LMICs compared with high-income countries. The participants do however have access to a strong public primary health program including well-developed antenatal services, a strong HIV prevention of mother-to-child transmission and a comprehensive childhood immunisation program.

Data on risk factors for wheezing from the antenatal period through 5 years were collected, including sociodemographic factors, nutrition, maternal physical and mental health, home environment, birth factors and breastfeeding, Table 1.2. Maternal mental health measures included measurements of depression, psychological distress, and intimate partner violence (IPV) antenatally and postnatally. Smoking was assessed by maternal self-report antenatally and postnatally. SES was assessed through a validated measure comprising 4 components: household income, asset ownership, household size and maternal education.

Active surveillance was used to confirm LRTI; all episodes were assessed by trained study staff and defined by WHO definitions, [23]. Episodes of LRTI or of wheezing which were <28 days apart were recorded as a single episode. At each LRTI or wheezing episode, a nasopharyngeal swab was obtained for multiplex real-time PCR using FTDResp33 (Fast-track Diagnostics, Luxembourg), identifying up to 33 organisms including respiratory viruses RSV, rhinovirus (RV), adenovirus (AV), influenza and parainfluenza virus, which has been previously associated with LRTI in this cohort, [24, 23].

Figure 1.1 depicts a comprehensive cohort flow diagram showcasing the primary respiratory outcomes of interest, namely childhood wheeze and lung function development. Cohort retention was high within the study, with 1015 (88.8%) of 1143 children retained at 1 year, 994 (87.0%) at 2 years, and 981 (85.8%) at 5 years. A unique aspect of the DCHS is the success rate and data granularity collected in early life. Lung function measurements were assessed from 6 weeks of age, with longitudinal measurements done annually. Data on childhood wheeze was collected at 14 different time points up to age five.



LRTI = Lower Respiratory Tract Infection; MBW=Multiple breath washout; AOS=Airwave oscillometry

* 632/1015 (62.3%) Children at year 1, 639/994 (64.3%) at year 2, and 825/981 (84.1%) at year 5 with lung function included in the analysis.

Figure 1.1: Cohort flow diagram.

Table 1.2: Definitions of child, maternal, and socio-economic characteristics from the DCHS

	Measurement used
Child characteristics	
Birth weight z-score	Z-scores were calculated using the Fenton newborn growth charts for preterm infants and WHO growth charts for term-born infants. Z-scores adjust for age and sex, using Fenton and WHO in combination to adjust for prematurity. Trained clinical staff recorded newborn birthweight at delivery, using a digital scale with a precision level of 10 g.
Preterm	Gestational age at birth <37 weeks. Gestational age was estimated using antenatal ultrasound measurements in the second trimester; if this was unavailable then symphysis-fundal height or maternal recall if the last menstrual period was used.
Late preterm	Gestational age at birth 34 to <37 weeks.
Lower Respiratory Tract Infection (LRTI)	World Health Organisation (WHO) criteria were used to define LRTI. (1) LRTI was diagnosed in children with a cough or difficulty breathing and age-specific tachypnoea (≥50 breaths per minute for children aged 2-12 months; > 40 for children 1-5 years) or if the child had lower chest wall indrawing. Severe LRTI was diagnosed in children younger than 2 months with tachypnoea (>60 breaths per minute) or lower chest wall indrawing, or in children of any age if the child had a general danger sign (cyanosis, unable to drink, seizures, or decreased level of consciousness). Episodes of LRTI less than 28 days apart were regarded as a single episode.
RSV-LRTI; RV-LRTI; AV-LRTI; Influenza (A or B or C)-LRTI; Parainfluenza (1 or 2 or 3)-LRTI	A LRTI episode with a positive PCR result for RSV, RV, AV, Influenza (A or B or C), or Parainfluenza (1 or 2 or 3) LRTI on a nasopharyngeal swab.
Current wheeze	Data on wheezing were obtained using validated questionnaires based on the ISAAC methodology at 14 scheduled visits from birth to 5 years or detected by auscultation by trained study staff. In addition, questionnaires were done at any unscheduled intercurrent illness episode.
Wheeze phenotypes	Never, early transient, late onset, and recurrent wheeze, as previously described by McCready et al. (2023), [25].
HIV exposed uninfected	A child born to a mother living with HIV but who was uninfected. HIV-exposed children were tested for HIV at 6 weeks (by PCR), 9 months (by PCR, ELISA or rapid antibody testing) and 18 months (by rapid antibody testing), as per provincial prevention of mother-to-child transmission guidelines.
Exclusive breastfeeding (at 6 weeks)	Maternal reported breastfeeding only at 6 weeks.
Maternal characteristics	
Smoking	Self-reported smoking was assessed using the Alcohol, Smoking and Substance Involvement Screening Test (ASSIST) antenatally, at birth and during the past three months postnatally. Postnatal smoking was defined at any time point from birth to 12 months, [26, 27].
Maternal asthma or allergy	Self-reported maternal asthma or allergy was assessed by direct interview antenatally.
Intimate partner violence (IPV)	IPV Questionnaire adapted from the WHO multi-country study was used to assess maternal physical, emotional, or sexual violence exposure antenatally and postnatally. Postnatal IPV collection was used from birth to 1-year. This included the 10-week, 6-month, and 12-month data points, [26].
Depression	The Edinburgh Postnatal Depression Scale (EPDS) was used to measure maternal depression antenatally and postnatally. 10 Questions were scored 0-3 and totalled. A cut-off value of 13 was used to separate the participants into above- or below-threshold groups. Postnatal depression was used from birth to 1-year. This included the 10-week, 6-month, and 12-month data points, [26].
Psychological distress	The Self-Reported Questionnaire 20-item (SRQ20) was used to measure maternal psychological distress antenatally and postnatally. Each item was measured 0-1; a cut-off value of 8 was used to distinguish an above- or below-threshold group, [26].
Socio-economic status (SES)	
Income	Average household income per month at maternal enrolment. Categories are Less than R1 000 (USD67), R1 000 (USD67) to R5 000 (USD336), and More than R5 000 (USD336).
Asset ownership	Asset ownership is a summed score of 13 different questions including access to electricity, tap or running water, domestic servant, flush toilet inside, built-in kitchen sink, an electric stove or hotplate, working telephone, at least one motor car or truck, motorcycle or scooter, a bicycle, shop at supermarkets, use any financial services, account at a retail store. The distribution across quartiles at maternal enrolment. The levels are Low, Low-Medium, Medium-High, and High.
Household size	The distribution across quartiles of household size (members) at maternal enrolment. The levels are Small (1-4), Small-Medium (4-5), Medium-Large (5-7), Large (7-18).
Maternal education	Highest maternal education level obtained at maternal enrolment at maternal enrolment. The levels are primary; some secondary; completed secondary education; and any tertiary education.

RSV = Respiratory syncytial virus; RV = Rhinovirus; AV = Adenovirus
HIV = Human immunodeficiency virus; PCR = polymerase chain reaction; ELISA = Enzyme-linked immunosorbent assay

1.2.4 Ethic approval

This study was approved by the Human Research Ethics Committee of the Faculty of Health Sciences, University of Cape Town, South Africa (HREC 401/2009). Mothers gave written informed consent at enrolment and were re-consented annually.

1.2.5 Data lock

For this thesis, a data lock was imposed in mid-2023 finalising the dataset ready for analysis, and then its state is kept constant. Before mid-2023, the data was cleaned and validated and considered ready for analysis. Subsequent changes and additions to the database are accordingly not reflected in the analysis.

1.3 Outline

This thesis is divided into five chapters. The initial two chapters lay the groundwork by introducing the study's context, research objectives, and literature review. Subsequently, chapters 3 and 4 deal with different aspects concerning the modelling and analysis of respiratory outcomes. Chapter 3 focuses on the identification of wheezing phenotypes while chapter 4 focuses on latent classes of lung function trajectories. Chapter 3 precedes chapter 4 since it hypothesised that one, or several, wheezing phenotypes in early life may be associated with impaired lung function development. Chapter 5 contains a concluding summary of the study and brings together the analyses of chapters 3 and 4 by investigating the association between wheezing phenotypes and lung function classes. While each chapter has its unique research objectives, the content is structured to enable integration and progression by allowing each chapter to build on the findings of the previous chapters.

1.4 Key contributions

While the clinical and statistical contributions and novelties of this thesis are mentioned throughout the results and discussion sections, this section provides a concise summary of key contributions.

Key contributions from Chapter 3: *The identification of wheezing phenotypes in early childhood using binary indicators.*

- Clinical: This study is the first to describe patterns of wheezing, associated risk factors, and the impact on lung function in a birth cohort from an LMIC. Spanning birth to age 5, it includes data from 14 time points with high cohort retention, capturing exposures that are quite different from those in HICs.
- Clinical: The study highlights the strong association between RSV-LRTI and recurrent wheezing, which continues through early childhood. This finding is important in the context of new preventive interventions for RSV.
- Clinical: A novel finding from the study is the association between recurrent wheezing and impaired lung function at 5 years, offering new insights into the long-term effects of wheezing phenotypes on respiratory health.
- Statistical: The study compares the statistical strengths and weaknesses of the PAM and LCA clustering methods in the DCHS context, contributing to a better understanding of how these approaches perform in longitudinal data setting.
- Statistical: The use of a repeated sampling approach for phenotype clustering refines the assessment of temporal stability in wheezing phenotypes, which improves reproducibility and reduces uncertainty.
- Statistical: The development of a tailored application and adaptation of a cross-validation (CV) framework is a key contribution, improving variable selection and ensuring stability in model estimates related to wheezing phenotypes.

Key contributions from Chapter 4: *Distinctive lung function trajectories from birth to 5-years in a South African birth cohort study.*

- Clinical: Even mild LRTIs occurring in the first year of life can have lasting effects, persisting to age 5. This is important because lung function at school entry is a strong predictor of the physiological peak achieved in early adulthood. Furthermore, the study provides a comprehensive description of lung development from birth to the preschool years, filling a key gap in current research.
- Clinical: The study linked different wheezing phenotypes with lung function trajectories, highlighting associations between wheeze onset and the development of respiratory resistance and airway obstruction.

- Clinical: The study extended previous findings by showing that HIV exposure in utero leads to sustained lung function impairments, particularly in children with mothers who had detectable HIV viral loads during pregnancy.
- Statistical: Demonstrated the flexibility of GAMLSS models in combination with an ITS approach in capturing complex growth patterns in height-adjusted lung function outcomes.
- Statistical: Applied LCMM to identify heterogeneous trajectories in lung function, revealing the existence of distinct latent classes with differing developmental patterns and showing how multivariate LCMM can uncover shared latent processes across multiple lung function dimensions.
- Statistical: Introduced a feature-based clustering approach using GAMLSS for trajectory modelling followed by PAM clustering, providing an alternative to LCMM solutions with dominant majority classes.

Chapter 2

Literature review

2.1 Aim and structure

In this literature review, the aim is to provide a comprehensive overview and progression of influential studies within the field of childhood wheezing and lung function development. This review will include an introduction to the studies, details of the statistical methodologies used, and a summary of the key findings. The methods employed by the various studies are briefly listed in this literature review chapter and then, where applicable, discussed in more detail in the methodology sections in chapters 3 and 4. Additionally, this literature review uncovers common patterns, differences, and areas where research is lacking in the current body of literature.

2.2 Childhood wheeze

Wheezing is common in early childhood and might be caused by a range of different pathophysiological mechanisms, [28]. The awareness of the heterogeneity within childhood asthma and wheezing and its varying response to treatment is steadily growing. This heterogeneity can be described in the form of phenotypes. In this thesis, *phenotypes* represent different observable profiles of wheeze, where each profile has its associated characteristics and risk factors.

Several main reasons why distinguishing phenotypes is important are summarised effectively by Spycher et al. (2010), [29]. Phenotypes can be used to enhance precision in research and clinical studies. Spycher et al. (2010) suggested that the inconsistent associations found in research studies between phenotypes and their risk factors may be explained by phenotypic variation. The ability to implement new prevention interventions is particularly important in LMICs. For example, identifying that a particular site or group develops recurrent wheezing at a higher rate due to tobacco smoke exposure allows for the implementation of smoking cessation programs in women of childbearing age. Phenotype-specific guidelines have been increasingly recognised by paediatric societies in preschool years, [30, 31]. New research projects present an opportunity to expand these existing

guidelines.

Birth cohort studies in high-income countries have described several wheezing phenotypes, with specific and shared risk factors, [15]. Howard et al. (2015) published a summary of 25 recent asthma cohorts that used LCA approaches to derive phenotypes. Of these 25 cohorts, 11 specifically investigated phenotypes based on childhood wheeze. Only one of these cohorts was based in an LMIC as classified by the Organisation for Economic Co-operation and Development (OECD), [32, 16].

In contrast to high-income countries, limited information exists on the patterns of childhood wheezing in LMICs, despite the high prevalence of wheezing illness and the large burden of risk factors and adverse environmental exposures associated with wheezing. This includes the high incidence and severity of early-life lower respiratory tract infections (LRTIs),[21] with respiratory syncytial virus (RSV), which is frequently associated with recurrent wheezing in high-income countries, being a predominant causal pathogen, [22]. Furthermore, high exposure to air pollution, tobacco smoke, psychosocial stressors, or poor living conditions might lead to different patterns of wheezing illness in LMICs compared with high-income countries. Therefore, understanding phenotypes of childhood wheezing illness in LMICs is important to understand the spectrum of disease, identify risk factors, and inform intervention strategies.

The lack of representation from LMICs may raise questions about the generalisability of findings from studies conducted predominantly in high-income countries. Phenotypes identified in an HIC context may not directly apply to LMIC populations with different living conditions and healthcare infrastructures. Including diverse populations in research ensures that the identified phenotypes are applicable and meaningful on a global scale.

The process of discovery has evolved from hypothesis-testing approaches, in which wheezing phenotypes are specified a priori based on clinical insights, to hypothesis-generating methodologies aiming to discover the latent structure of patterns of wheezing onset, remission and recurrence within large data sets, which use a variety of unbiased machine learning and statistical methods. Early approaches to wheezing phenotype discovery were simplistic by clustering children together based on a few simple indicators (triggers) of wheezing, such as the frequency of wheeze episodes. The Melbourne Asthma Study, conducted from 1964 to 1999, was one of the earliest respiratory studies which applied hypothesis-testing approaches to wheeze phenotype discovery in children, [33]. The Melbourne children were divided into four distinct groups based on specific criteria measured by simple indicators: (1) A group of children who never had reported wheeze, (2) who had wheezed less than five times and had bronchitis association, (3) wheezed five times or more and had bronchitis association, and (4) an asthma group which contained children who had at least one episode of wheeze, and the wheezing was not linked to respiratory infection.

The Tucson Children's Respiratory Study, which began in 1980, is another early study which characterised wheezing phenotypes by patterns of wheezing at 6 years of age. The phenotypes included

never wheezing children (n=425, 51.5%), transient early wheezing (wheezing by age 3 years, but not after, n=164, 19.9%), persistent wheezing (wheezing at both ages 0-6 years, n=113, 13.7%) and late-onset wheezing (wheezing onset between 4 and 6 years, n=124, 15%), [34]. Subsequently, the Avon Longitudinal Study of Parents and Children (ALSPAC) birth cohort identified generally similar phenotypes using latent class analysis (LCA) approaches, [35].

The presence of wheezing phenotypes was investigated in the ALSPAC, a birth cohort study established in 1991 in Avon, UK with 14062 live births. The study utilised LCA to identify distinct phenotypes by analysing wheezing patterns recorded at 14-time points from birth to 16.5 years. The authors described 6 different phenotypes, never/infrequent, preschool-onset remitting, mid-childhood onset remitting, school-age onset persisting, late-childhood onset persisting, and continuous wheeze, [36, 37].

The PIAMA Study (Prevention and Incidence of Asthma and Mite Allergy), conducted by Dutch authors, focused on evaluating 3963 infants, with a subset of 2728 individuals undergoing assessment, spanning from 0 to 8 years. Five phenotypes were identified using longitudinal latent class analyses (LLCA): never/infrequent wheeze (75%), transient early wheeze (16.7%), intermediate onset wheeze (3.1%), persistent wheeze (3.5%) and late onset wheeze (1.7%), [38].

Savenije et. al (2011) analysed and compared the phenotypes between the PIAMA Study and the ALSPAC study using LLCA at 8 comparative time points,[39]. The wheezing phenotypes were found to be highly comparable in both birth-cohort studies, although the prolonged early phenotype identified in ALSPAC was not replicated in the PIAMA cohort.

The ELFE cohort (French Longitudinal Study of Children) is a birth-cohort study which tracked 18041 newborns at 2 and 12 months of age. The authors assessed the presence of wheezing at both visits and described four wheezing phenotypes. The phenotypes included non-wheezers, who made up 7.2% of the cohort, 21.1% had intermittent wheezing (episode of wheezing at two months), 2.4% had persistent wheezing (wheezing episodes at 2 months and 1 year), and 18.3% experienced incident wheezing (when the first episode of wheezing occurred after the age of two months), [40].

Lodge et. al (2014) investigated the existence of phenotypes in the Melbourne Atopy Cohort Study (MACS), which followed 620 children at highrisk for allergy from birth to the age of 7, as well as subsequent follow-ups at the age of 12, [41]. The study employed LCA to identify distinct patterns of wheezing. Results revealed five wheeze phenotypes. The phenotypes included never/infrequent wheeze 42.7%, early transient wheeze 27.5% (appearing in the first twelve months of life and resolving permanently around three years), early persistent wheeze 5.7% (manifesting in the first 6 months of life), intermediate-onset wheeze 20.7% (emerging around 18 months), and late-onset wheeze 3.5% (occurring after the age of four).

Oksel et. al (2019) pooled data from 5 UK population-based birth cohorts: ALSPAC [36], Ashford (ASHFORD) [42], and Isle of Wight (IOW) [43] cohorts; Manchester Asthma and Allergy Study (MAAS) [44]; and the Aberdeen Study of Eczema and Asthma to Observe the Effects of Nutrition

(SEATON) [45]. After which LCA was used to identify 5 wheezing phenotypes from 7719 participants: never/infrequent wheezing (52.1%), early onset preschool remitting (23.9%), early onset mid-childhood remitting (9%), persistent (7.9%), and late-onset wheezing (7.1%).

The studies on wheezing phenotypes can be broadly divided into two categories: those that utilised a priori-defined phenotypes and those that employed data-driven approaches, such as LCA. The earlier studies, such as the Melbourne Asthma Study and the Tucson Children's Respiratory Study, used hypothesis-testing methods with predefined phenotypes based on clinical insights and specific criteria like wheezing frequency and patterns. In contrast, more recent studies, including the ALSPAC and PIAMA cohorts, have adopted hypothesis-generating techniques that utilise LCA to identify phenotypes based on data-driven patterns of wheezing onset, persistence, and remission. There is substantial agreement between a priori defined and data-driven approaches in identifying broad wheezing phenotypes, such as early transient, late onset, and a persistent/intermittent phenotype, across studies. This consistency suggests that while predefined phenotypes were simpler, they captured fundamental patterns that are echoed in more sophisticated, data-driven classifications. However, LCA and similar methods often identified additional nuanced phenotypes, such as intermediate onset or remitting types, that were not captured by earlier studies. Therefore, data-driven methods provide a more granular and detailed understanding of wheezing trajectories.

Up to this point, the more recent birth cohorts which have been introduced have mostly employed model-based approaches, such as LCA or variations thereof, to identify wheezing phenotypes. Haider et. al (2022) published a robust and novel data-driven method by which repeated observations of wheezing episodes are used to generate multidimensional indicators of wheezing spells. These indicators encompassed factors such as duration, age of onset, and the degree of recurrence. This aligns with the recent Global Initiative for Asthma (GINA) guidelines which emphasise the importance of considering the frequency and severity of wheezing episodes, as well as the temporal pattern of symptoms, in the clinical diagnosis of asthma in children under 5 years, [46]. Her method provides a distance-based solution to phenotype identification as an appropriate distance matrix is calculated using the indicators before being used as input to a PAM algorithm, [47]. The distance-based approach summarises the longitudinal data points using cross-sectional indicators of the data. An analysis involving 7719 participants pooled from 5 UK population-based birth cohorts, which were introduced above by Oksel et. al (2019), unveiled five distinct spell-based wheeze phenotypes. The phenotypes included never wheezing (54.1%), early-transient (ETW) (23.7%), late-onset (LOW) (6.9%), persistent (PEW) (8.3%), and a newly discovered phenotype among the cohorts termed intermittent wheeze (6.9%).

Data-driven approaches have mostly used latent class models such as the LCA[48, 49, 50] to describe the within-individual and between-individual variability over a specified interval. The number of reported latent class analysis phenotypes varied by study, [51]. Furthermore, although latent class analysis phenotypes in different studies are usually designated by the same name, they often have different temporal trajectories (onset, duration, and recurrence), distributions within a population, and associated risk factors, [15, 51]. This reflects the complexity and heterogeneity within childhood

wheezing manifestations. The nuanced exploration of these phenotypes underscores the importance of discerning subtle differences, even when they share common names across studies, to enhance the understanding of the diverse trajectories in childhood wheezing. The subtle differences in nomenclature may reflect variations in study design, cohort characteristics, and analytical methods. This underscores the necessity for careful consideration and cross-referencing when interpreting and comparing wheezing phenotypes across different research studies.

The use of LCA to identify wheeze phenotypes is typically done using data that captures wheeze patterns through binary variables indicating the presence or absence of wheezing in children. However, recent studies have shown that employing binary wheezing data for LCA might result in the imprecise classification of individuals, [49, 51, 52]. It is further reported that under the LCA approach, children with identical wheezing patterns may be assigned to different phenotypes[51]; this is most common with intermediate phenotypes, focused on the timing of wheezing, such as early transient or late onset phenotypes. Whereas, there is a high degree of stability in the recurrent wheeze phenotype and wheeze-negative phenotype.

It is hypothesised that wheeze phenotypes in early childhood and their risk factors in LMICs (in particular early-life LRTI) differ from those in high-income countries and that the causes of early-life LRTI will affect the presentation, development, and recurrence of wheezing.

PubMed was searched for articles published in English from 1990 until 1 January 2025. A suitable keyword search[1] revealed a rise in the number of publications containing the terms phenotype, asthma or wheeze and paediatric or child, progressing from 101 publications in the 1990s to 796 publications between 2000 and 2009, and further escalated to 1821 publications between 2010 and 2019. While 1301 publications were observed between the start of 2020 to 1 January 2025.

To the best of current knowledge, the DCHS is the first birth cohort to describe patterns of wheezing, associated risk factors, and impact on lung function in an LMIC, where exposures differ from those in high-income countries in previous studies.

2.3 Lung function development in early life

Technological advances have allowed non-invasive measurements of lung function from infancy to adulthood, allowing the investigation of lung function development in relation to early life events and exposures. These measurements are possible even in lower-resourced communities where children are particularly at risk of poor respiratory health, [53].

Assessing lung function during early childhood holds significant importance, given the growth and development of the respiratory system. It is common for children in this stage to experience recurrent coughs and wheezing. While a majority of children eventually overcome their symptoms, some

[1] Articles must contain the keyword "phenotype", and the keywords "wheeze" or "asthma" and "pediatric" or "child".

will continue to suffer from persistent asthma well into adulthood, [34]. Multiple studies have investigated which factors would have an impact on the growth and development of lung function in early life and adolescence. Early life factors, such as extreme tobacco smoke exposure and premature birth, can directly impact early airway function, resulting in poor respiratory health. These impacts may persist, leading to compromised airway function and elevating the risks of wheezing and asthma, [54]. Additional factors which were identified include breastfeeding [55, 56], prenatal stress[57], allergies and atopy[58], maternal perinatal BMI[59], childhood BMI[60], and exposure to air pollution, [61]. Disadvantaged socioeconomic circumstances in early life have also been shown to impact lung function, [62]. While some of these studies have concentrated on examining an individual or a small set of characteristics, it is hypothesised that the fluctuations in lung function stem from the combined influence of multiple factors working simultaneously, [63].

The same risk factors are not always identified across studies. While some risk factors like extreme tobacco smoke exposure and premature birth are consistently highlighted as detrimental to lung health, other studies have identified a broader range of risk factors. For instance, factors such as prenatal stress, exposure to pollutants[61] and allergens in early childhood have been associated with altered lung function. The discrepancies in identified risk factors across studies can be attributed to various factors including study design, population demographics, and the specific outcomes measured. Different studies may focus on distinct aspects of lung function or target varying age groups, leading to diverse findings.

The DCHS has been central in investigating the determinants of early lung function in African infants. Infants enrolled in this South African birth cohort had their lung function assessed at 6 weeks and 1 year, [64, 65]. Research from the DCHS has shown that lower respiratory tract illness in early life impairs lung function at 1 year, independent of baseline lung function, emphasising the importance of understanding early-life exposures to childhood lung disease, [65]. Factors such as infant size, sex, maternal smoking, alcohol exposure, HIV exposure, and household benzene have been linked to altered early lung function, [64, 65].

The Tucson Children's study, the American birth cohort introduced earlier in this chapter is an example of a cohort from a HIC which investigated both wheezing phenotypes and lung function development in early life. The Tucson Children's study showed pneumonia in the first 3 years of life to be associated with reduced spirometry in childhood that persisted to early adulthood, [66]. However, most children globally live in low or LMIC settings and are particularly at risk of impairment in lung function given the high burden of early life risk factors, especially high incidence of severe LRTI,[67] exposure to polluted air and social disadvantage. Identifying which factors impair healthy lung development in early life in LMICs is important to specify targets for strengthening health throughout an individual's lifespan.

Research by He et al. (2023) addresses the association of early life and pubertal growth with lung function development (forced vital capacity (FVC), forced expiratory volume in the first second (FEV1), FEV1/FVC) and wheezing in a non-Western setting, [68]. In a representative Chinese

birth cohort, they reported that rapid weight growth in early life was linked to irregular differences in lung function development and a higher risk of wheezing and asthma. Thereby, they linked a child's size to both main respiratory outcomes of interest (wheezing and lung function) in this PhD thesis. They analysed lung function z-scores which were calculated using the appropriate South East Asian equations from the Global Lung Function Initiative (GLI) to adjust for age, sex, and height, [69]. History of wheezing was self-reported and defined as ever-wheezed at approximately 17.5 years. Larger size (kg, cm, and BMI) was positively associated with childhood wheezing. This association was strongest in early life (0-2 years). The work done by He et al. (2023) suggests the importance of considering the impact of weight growth velocity and size, especially in infancy, when modelling lung function development.

Age- and size-related reference ranges for spirometry through childhood and adulthood have been studied. In the study by Cole et al (2009), the LMS method is extended to references that depend on body size as well as age, by exploiting the flexibility of the generalised additive models for location, scale and shape (GAMLSS) approach, [70]. The paper discusses the development of models for lung function variables in relation to height and age, specifically forced expiratory volume in FEV1, forced vital capacity (FVC), and FEV1/FVC. It explores the multiplicative association between spirometry, height, and age. Cole et al (2009) reported that variability in lung function measurements depends on age and to a smaller degree on height. Understanding these associations can aid in the interpretation of spirometry results.

Typically, longitudinal data in health databases are analysed through mixed effect models, which estimates the average trajectory of the cohort in terms of fixed effects and the variation of the individual trajectories around the average in terms of subject-specific random effects. An application of this approach can be found in the initial analyses presented in the earlier publications from the DCHS, where lung function outcomes were modelled from birth to 2 years of age using mixed-effects linear regression modelling with random subject effects, [24]. The focus can be extended from a single average trajectory towards investigating the presence of multiple trajectories, [71]. Resulting in the identification and characterisation of latent subgroups of individuals exhibiting a shared profile throughout the observed study period. Similarly to the review on childhood wheezing, latent subgroups identified using lung functional data points can also be referred to as phenotypes or classes[71], this thesis will refer to them as latent classes or classes. The exploration of latent classes of trajectories can be extended by investigating their relationship with explanatory variables. This analysis would assist in identifying key determinants that contribute to the heterogeneity observed within latent classes.

For clinicians, gaining insight into the effectiveness of a lung function measure in distinguishing between health and disease is useful. If a disease is present in a child, longitudinal measurements could assist in effectively tracking the progression of a disease. However, lung function measures in both healthy individuals and those experiencing respiratory symptoms or diseases frequently exhibit significant overlap, to the extent that a normal lung function measurement does not definitively rule out the presence of disease, [10]. Given this recognised overlap, latent classes of lung function

trajectories may not exclusively indicate the presence of a specific disease, but can also represent a natural trajectory observed in disease-free infants.

In 1977, Fletcher and Peto, based on their 7-year epidemiological study of male postal workers in London, proposed the hypothesis that individuals in a population may follow distinct trajectories of spirometric decline, influenced by the presence of specific risk factors such as smoking, [72]. Since then, multiple studies involving population-based cohorts have investigated whether lung function follows predictable patterns of change over time. These studies vary in their methodological approaches to identifying lung function trajectories, the characteristics of the cohorts, the age range of participants, and the duration of observation.

Berry et al. (2016) used a latent class linear mixed effects model, adjusting for sex, to analyse longitudinal measurements of the ratio of FEV1 to FVC (FEV1/FVC) from the Tucson Children's Respiratory Study, [73]. They identified two distinct trajectories: a low trajectory (9.3%, n = 56) and a normal trajectory (90.7%, n = 543). Their findings suggest that persistently low lung function may be to some extent established at birth.

The Childhood Asthma Management Program (CAMP) identified 4 trajectories of FEV1 among children with childhood asthma, namely, normal growth, reduced growth, normal growth (early decline), and reduced growth (early decline), [74]. Pattern classification was conducted subjectively by experts based on the overall shape of participants' FEV1 curve over time, using NHANES III reference curves as rough guidelines.

In the Manchester Asthma and Allergy Study (MAAS) from 5 to 16 years and the Avon Longitudinal Study of Parents and Children (ALSPAC) from ages 8 to 24 years, Belgrave et al. (2018) used latent profile modelling to identify 4 trajectories of FEV1 labelled as persistently high, normal, below average, and persistently low, while assuming that FEV1 profiles change linearly, [75]. Of interest, Belgrave et al. (2018) reported that the persistently low trajectory was associated with persistent wheezing and asthma.

Kim et al. (2018) identified 5 distinct FEV1 types in non-smoking asthmatics drawn from a Korean Asthma cohort using a k-means clustering approach for longitudinal data, [76].

Karmaus et al. (2019) applied a finite mixture model to longitudinal data of FVC, FEV1, the FEV1/FVC ratio, and FEF25-75 from the Isle of Wight (IoW) birth cohort at ages 10, 18, and 26 years, [77]. They identified two distinct trajectories for males and females separately: a low trajectory and a high trajectory.

Horsley et al (2022) applied latent class growth analysis (LCGA) by using the LCMM package in R to identify distinct trajectories of LCI data from the multiple breath washout (MBW) test, [78]. They analysed individuals diagnosed with mild cystic fibrosis (CF) lung disease, to determine what temporal variations occur in the LCI, and which factors contribute to these changes over time. They found 4 distinct clusters, based on the trajectory of LCI. These were a large stable near normal

(n=58, 72%), low LCI rising over time (n=8, 10%), elevated LCI falling over time (n=7, 8%), and high LCI increasing over time (n=8, 10%).

Sanna et al. (2022) applied finite mixture modelling to characterise lung function trajectories over 16 years (6- to 22-years) from childhood to young adulthood, using a cohort from the Raine Study, a prospective pregnancy cohort from King Edward Memorial Hospital in Perth, Western Australia, [79]. Four lung function trajectories of FEV1, FVC, and FEV1/FVC (z-scores) were identified in the study.

More recently, Granell et al. (2024) analysed data from three population-based birth cohorts: ALSPAC, IOW, and MAAS, [80]. A total of 6,377 participants who completed spirometry on at least two occasions were included in the analysis, with ages 8-26 across all three cohorts. Latent profile modelling was used to derive trajectory classes based on the development of FEV1 to FVC ratio over time. Where, two-level random intercept regression models were used to assign participants to their most likely trajectory profile, comparing models for goodness-of-fit using the BIC. The study by Granell et al. (2024) identified four distinct FEV1 to FVC ratio trajectories: Above Average (49.5% of participants), Average (38.3%), Below Average (10.6%), and Persistently Low (1.7%).

The clustering approaches used in various studies to analyse lung function trajectories show both consistency and inconsistency. Consistency is seen in the overarching goal of identifying distinct patterns of lung function development or decline over time and the use of longitudinal data. Inconsistencies arise from differences in methodological approaches, cohort characteristics, age ranges, and the number and nature of identified trajectories. Studies vary in their methods such as latent class linear mixed effects modelling, subjective pattern classification, and finite mixture modelling. Cohorts differ in terms of populations studied, age ranges, and observation durations. The number of identified trajectories varies across studies, with differences in the nature of these trajectories like persistently low, normal, or reduced growth patterns.

A further clustering approach was applied by Augustin et al. (2018) who identified 7 distinct lung function clusters through the application of self-organising maps (SOM), [81]. SOMs can be extended to solve clustering problems that involve multidimensional temporal trajectories. Although not conducted in the medical research field, Ling and Delmelle (2016) demonstrated a two-layered approach to classify neighbourhoods according to their multidimensional temporal trajectories[82]; the theoretical approach underlying the application by Ling and Delmelle (2016) is transferable to solve clustering problems within the lung function trajectories domain. The application of SOM within the context of lung function trajectories and their potential as a robust substitute for more common clustering approaches, such as latent class mixed modelling, remains largely less studied.

The comprehensive lung function measures from birth through early childhood in the DCHS are unique, especially in an LMIC context, and provide particular advantages. For instance, LCI has an advantage over conventional FEV1 measurements. These include greater sensitivity in identifying airway obstruction, improved tolerance in young children, and a consistently stable upper limit of

normal, [11, 78]. Specific lung function measures are recognised as valuable endpoints in clinical analysis. LCI, for example, is known as an important endpoint for individuals with cystic fibrosis, [78].

Critically assessing the application of these clustering approaches is important. Model-based techniques such as latent class mixed modelling (LCMM), LCA, and partitioning techniques such as k-means are increasingly popular methods in the field of respiratory health, but there are reservations about the replicability, reliability, and reporting of findings, [83, 84]. A common definition of replicability is that the classes must be reproducible across multiple samplings from the same population in question, [85]. Clustering solutions that result in higher replicability would be considered more valid, [85].

The application of longitudinal clustering is increasingly prevalent. PubMed was searched for articles published in English from 1990 until 1 January 2025. A suitable keyword search[2] revealed a notable rise in publications related to longitudinal clustering over time, progressing from 143 publications in the 1990s to 442 publications between 2000 and 2009, and further escalated to 2272 publications between 2010 and 2019. While 2278 publications were observed between the start of 2020 to 1 January 2025. This upward trend highlights the growing recognition of longitudinal clustering in defining developmental trajectories and addressing heterogeneity in pediatric research.

[2]Articles must contain the keyword "longitudinal", and one of the keywords "mixture", "latent-class", "latent-class", "clustering", or "group-based", along with one of the keywords "pediatric" or "child"

Chapter 3

The identification of wheezing phenotypes in early childhood using binary indicators

3.1 Introduction

3.1.1 Chapter aims

The developmental trajectories of childhood wheezing in LMICs have not been thoroughly described. This chapter aims to establish longitudinal wheeze phenotypes by analysing longitudinal binary observations of self-reported wheeze recorded from birth to 5 years of age in the DCHS.

The work introduced in chapter 3 addresses scientific questions from both clinical and statistical standpoints. Clinically, this analysis aims to identify and describe longitudinal patterns of wheezing, associated risk factors, and the impact on lung function in an LMIC, where exposures differ from those in high-income countries in previous studies. To the best of current knowledge, this is the first birth cohort to do so. The identified phenotype structure is further compared with a similar analysis conducted on data from the ALSPAC study, a UK birth cohort study, [36]. Comparing phenotype structures between studies helps validate findings, enhance generalisability, understand variability, and refine phenotypic definitions.

From a statistical perspective, chapter 3 focuses on the critical review and application of clustering binary wheezing responses observed in a clinical setting. Different distance-based and model-based approaches are investigated to identify latent profiles through a critical review, application, and comparison. In doing so, we compare a newer and more innovative distance-based technique to the well-established model-based LCA approach. The performance of these approaches is compared using (1) the number of phenotypes detected, (2) phenotype reproducibility and stability, assessed through a repeated sampling approach to examine the impact of sample size and input data order on

clustering results, and (3) the correspondence between the indicators used to derive the phenotypes and clinical knowledge and interpretation. The analysis further aims to investigate the associations of wheezing phenotypes with early-life risk factors, including specific viral LRTI and baseline lung function at 6 weeks. Additionally, it sought to explore the association of different wheeze phenotypes with lung function at 5 years, providing valuable insights into the impact of these factors on the development and progression of wheezing phenotypes in early childhood.

3.2 Data sources and definition of variables

The predominant definition of wheezing comes from the International Study of Asthma and Allergies in Childhood (ISAAC). In the DCHS, childhood wheezing was assessed using ISAAC questionnaires by trained study staff at a study visit or during an intercurrent illness, [24]. Current wheezing was defined as a positive response to the question "Has your child had wheezing or whistling in the chest since the last visit?" at each follow-up. Validated questionnaires were completed at 14 scheduled visits from birth until 5 years (six visits in the first 12 months, and at 6-monthly intervals thereafter) as shown in Figure 3.1. Collecting data from multiple visits across various ages is crucial due to the longitudinal nature of childhood wheezing, which is influenced by age-dependent risk factors.

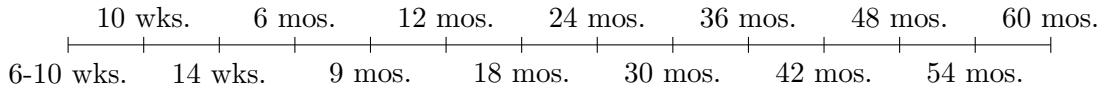


Figure 3.1: Scheduled Visits

This analysis included 950 (83%) of 1143 children with complete data on wheezing across 14 time points to 5 years with high cohort retention. At least one episode of wheezing, defined through self-report, occurred in 470 (49%) of 950 children; 256 (27%) had one, 112 (12%) had two, and 102 (11%) had three or more episodes. The median age of the first wheezing episode was 8.3 months (IQR 3.3-16.8 months). The prevalence of current wheeze across the 14 time points peaked at 13.3% at 6 months and declined to 1.5% by 5 years, Figure 3.2. Within the DCHS, there is a notably rapid decline in wheeze (1.5% at age 60 months), especially compared with the ALSPAC cohort (18% at 54 months) which was studied in the UK, [48]. The decline in wheezing may reflect lower exposure to viral-induced wheezing in older children, less exposure to other risk factors such as indoor tobacco smoke or growth in calibre of the airways resulting in less airway obstruction and symptomatic wheezing. Lack of reporting in children in the later pre-school years may also be a possible factor.

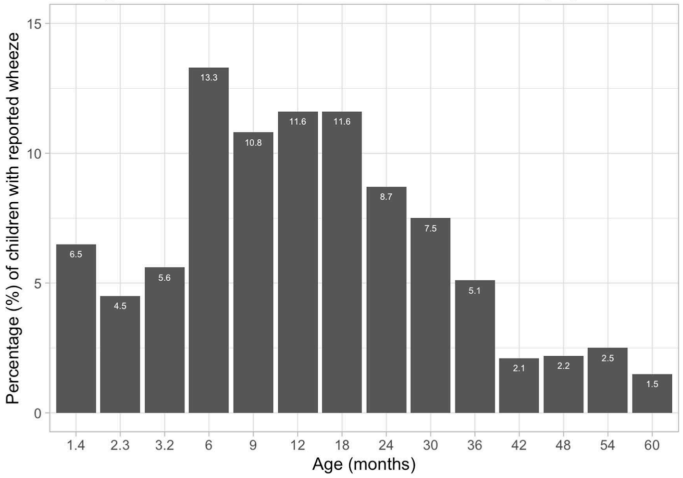


Figure 3.2: Point prevalence of current wheeze in children in the DCHS by age.

3.3 Longitudinal visualisations of wheezing phenotypes

In a longitudinal setting, visualisations are particularly useful for indicating how clusters vary over time. This study employs two visualisation techniques in this chapter while investigating profiles of wheeze, these are heatmaps and disease prevalence plots.

Heatmaps represent a data matrix for a variable of interest or dissimilarity matrix through a system of colour-coding. In the context of wheezing phenotypes, heatmaps consist of rows that represent binary data for subjects and columns that represent a primary outcome variable of interest measured longitudinally. The interpretability of a heatmap depends on the order of the rows. If an inherent ordering does not exist, an ordering should be selected that best displays patterns in the data, such as sorting based on the total number of wheezing episodes per subject from largest to smallest.

Prevalence is the proportion of disease-positive cases in the cohort at a given time. Later in this chapter, disease prevalence plots are used to display the observed prevalence of a disease in different wheezing classes across multiple study visits. Mathematically, prevalence at a visit is defined as:

$$\text{Prevalence} = \frac{\text{Number of disease-positive cases}}{\text{Class size}}, \tag{3.1}$$

expressed as a percentage. The prevalence at a given study visit signifies the proportion of subjects within that class exhibiting a positive outcome for the disease. These plots can reveal patterns, trends, similarities, and differences between clusters that may not be evident in summary tables. The x-axis typically represents an independent variable, such as time, visit number, or age group, while the y-axis shows the dependent variable, which in a clinical setting, is usually the proportion

of subjects with the disease of interest.

Both visualisation techniques, when used together, will enhance the interpretation and definition of phenotypic characteristics.

3.4 Distance-based clustering

The following methodology section of this chapter is divided into two main parts. The first part outlines the methodology for distance-based clustering, while the second part discusses the methodology for model-based clustering. The clustering methods include a partition around medoids (PAM) approach tailored for longitudinal data, which is compared to the results obtained using latent class analysis (LCA) and LCA with random effects. This will highlight the strengths and limitations of each method in handling data complexities within the DCHS.

Justifying the choice of the methods, LCA is widely regarded as the gold standard for deriving wheezing phenotypes, [16]. However, PAM provides a newer, non-parametric alternative that is well-suited for longitudinal data. This aligns with the methodology used in the ALSPAC cohort, allowing for a more direct comparison of wheezing phenotypes between a HIC cohort (ALSPAC) and one from a LMIC (DCHS).

3.4.1 Distance measures

In the context of distance-based clustering, distance matrices play an important role in identifying and grouping similar subjects based on their characteristics. Accurately measuring differences between individuals is crucial, particularly when studying wheezing or asthma phenotypes, as it ensures that clusters genuinely represent similarities in phenotypic expressions. In practice, distance matrices can be calculated from continuous variables, dichotomous variables, or mixed data (categorical or continuous).

Clustering algorithms group similar data points together based on their distances using a distance matrix - a symmetrical matrix in which subjects (objects) that are similar will have a smaller distance, where the distance between pairs is a measure of association. Therefore, given m subjects (objects), $\frac{1}{2}m(m-1)$ different pairs of distances will be observed. Next, the defined distance matrix would be used as input to the partitioning (or hierarchical) clustering algorithm. While using distance-based clustering, no assumptions are made concerning the number of groups or the underlying structure of the data.

The Euclidean distance and the Manhattan distance are two widely applied dissimilarity measures. Both distance measures are distinctive cases of the Minkowski distance, [86].

Denote the observed value of the t^{th} variable of the object i by x_{it}, $t = 1, ..., T$. The T-dimensional vector $\mathbf{x}_i = [x_{i1}, ..., x_{iT}]$ represents data for each object over T variables. Denote the dissimilarity between objects by $d(\cdot, \cdot)$.

The Euclidean distance between two objects i and i' is defined by

$$d(\mathbf{x}_i, \mathbf{x}_{i'}) = \sqrt{\sum_{t=1}^{T} (x_{it} - x_{i't})^2} \ . \tag{3.2}$$

The Manhattan distance between two objects i and i' is defined by

$$d(\mathbf{x}_i, \mathbf{x}_{i'}) = \sum_{t=1}^{T} |x_{it} - x_{i't}| \ . \tag{3.3}$$

Different dissimilarity functions are available to accommodate the various scales of measurement for variables, including nominal (e.g., gender, ethnicity), ordinal (e.g., socioeconomic status: "low", "low-middle", "middle-high", and "high'), interval (e.g., temperature in Fahrenheit scale), and ratio (e.g., ratio of time of peak total expiratory flow to time of expiration, t_{PTEF}/t_E). Where nominal and ordinal types represent categorical data and interval and ratio types represent numerical or quantitative data. One of the primary assumptions underlying these standard methods for calculating dissimilarities is that the variables used as input into Equations 3.2 and 3.3 are continuous, [87]. Quite often, situations exist where various variables of interest were measured on different scales which requires a distance matrix capable of coping with mixed type variables (categorical and numerical types). For example, this situation holds true in the DCHS, where wheeze can be recorded both as the total number of wheezing episodes a child experienced over the observational period and as a categorical variable, such as no wheeze, single wheeze, or intermittent wheeze.

Distance measures for dichotomous variables

Many cases in clinical research exist where the outcome of interest is the occurrence of an event over a fixed period which is presented in a binary matrix. For example, if the focus is on the developmental patterns of childhood wheezing, the outcome variable has two categories: wheezing present (yes) or wheezing absent (no). As before, cluster analysis with binary data starts by calculating distances between observations based on the observed variables and then applying the appropriate clustering algorithms to these distances.

In the presence of binary data, the Simple Matching Coefficient (SMC)[88] and the Jaccard index[89] are well-known methods used to calculate the dissimilarity (or similarity) between two objects. The difference is that the SMC includes both positive (a) and negative matches (d), while the Jaccard index only includes positive matches (a). The Jaccard index removes the possibility that 2 row vectors would be considered similar because neither contains many of the attributes of interest, [89]. These are defined in Equations 3.4 and 3.5.

$$\text{SMC} = \frac{a + d}{a + b + c + d} \ . \tag{3.4}$$

$$\text{Jaccard index} = \frac{a}{a+b+c} \ . \tag{3.5}$$

Consider two sets of binary variables, each consisting of p values that are either 1 (presence) or 0 (absence). In Equations 3.4 and 3.5, a represents the number of cases where both sets have a value of 1 (positive matches), b and c represent the number of cases where one set has a value of 1 and the other has a value of 0, and d represents the number of cases where both sets have a value of 0 (negative matches). The total number of cases (or comparisons) is $a+b+c+d$.

The underlying logic says that the similarity of two individuals is defined based on the shared pattern in the presence or absence of the binary attribute of interest, [90]. In both equations 3.4 and 3.5, the larger the value is, the higher the degree of similarity between objects.

It is recognised that a large number of similarity and dissimilarity measures between binary vectors are discussed in the literature in addition to the SMC and the Jaccard index. A summary of 79 binary similarity and dissimilarity measures was presented by Wijaya et al. (2016), [91].

Distance measures for mixed data

A plurality of distance measures are computed using either categorical or continuous data only. However, in practical settings, the datasets are often described by a combination of continuous and categorical variables. Data sets such as these are referred to as *mixed-type* data.

One solution to deal with a mixed-type data set is to blend the variables into a dataset that contains the same data type throughout. This would require either re-coding the categorical variables as continuous variables or re-coding the continuous variables as categorical variables.

Alternately, a distance measure that can be applied directly to mixed-type data can be computed, as opposed to re-coding a subset group of variables. Gower's distance measure is one of the best-known measures for mixed-type data, [92].

As before, let $\mathbf{X}$ represent an $n \times p$ matrix, but now $\mathbf{X}$ contains both continuous and categorical variables measured simultaneously. Using Gower's similarity measure, a pair of objects $(x_i, x_{i'})$ can be compared on a variable k:

$$g(x_i, x_{i'}) = \frac{\sum_{k=1}^{p} w_k(x_i, x_{i'}) s_k(x_i, x_{i'})}{\sum_{k=1}^{p} w_k(x_i, x_{i'})}, \tag{3.6}$$

where $w_k(x_i, x_{i'})$ is the weight of variable k for the pair of objects $(x_i, x_{i'})$, with $i, i' \in \{1, 2, ..., n\}$ and $i \neq i'$. Next, $s_k(x_i, x_{i'})$ represents the similarity between a pair of objects $(x_i, x_{i'})$ compared on a variable k. If the value for both objects is non-missing the value of the variable weights is set equal to $w_k(x_i, x_{i'}) = 1$. In comparison, when a variable's value for one or both objects is missing, the corresponding variable weight is set as $w_k(x_i, x_{i'}) = 0$. Differential weights w_k could be decided on *a priori* and then be specified in equation 3.6 to raise the importance of specific

variables that are considered more relevant. The complement of $g(x_i, x_{i'})$ is the distance measure, $dist(x_i, x_{i'}) = 1 - g(x_i, x_{i'})$.

The scores for $s_k(x_i, x_{i'})$ are assigned as:

- For binary variables, $s_k(x_i, x_{i'}) = 1$, if both x_{ik} and $x_{i'k}$ are 1 (positive match) or both x_{ik} and $x_{i'k}$ are 0 (negative match). In cases of asymmetric binary outcomes of variable k where the negative (0) outcomes are less important, the corresponding variable weight is set as $w_k(x_i, x_{i'}) = 0$. Categorical variables are dealt with similarly as a generalisation of the binary case.
- For continuous variables a ranged-normalised distance is considered, $s_k(x_i, x_{i'}) = \frac{|x_{ik} - x_{i'k}|}{R_k}$, where R_k is the range of variable k in the sample or population. $s_k(x_i, x_{i'})$ will range between 0 and 1, where a value approaching 1 represents two similar objects and 0 represents two objects which differ perfectly.

Subsequent modifications have been proposed over the years on Gower's (1971) measure[92], such as Wishart[93] (2003) who proposed a modification to the continuous variable component. Instead of a range like Gower, it incorporates the variance of continuous variables in the distance calculation. The Wishart (2003) measure is shown in Equation 3.7.

$$d_{ij} = \sqrt{\sum_{l=1}^{p} \omega_{ijl} \left(\frac{x_{il} - x_{jl}}{\delta_{ijl}} \right)^2}, \tag{3.7}$$

where $\delta_{ijl} = s_l$ when the variable is numerical, and $\delta_{ijl} \in \{0, 1\}$ when the variable is binary or categorical. The literature demonstrates the recent utilisation of the Wishart distance matrix in the characterisation of wheezing phenotypes, [47]. In addition to Wishart (2003), there are other available distances for mixed variable data such as Huang[94] (1997), Podani[95] (2000), Harikumar and PVs distance[96] (2015), and Ahmad & dey's (2007) distance[97], which is a slight modification on the proposed method by Huang[94] (1997).

3.4.2 Partitioning methods

K-Means approach

The k-means clustering algorithm has proven effective in identifying asthma phenotypes, as demonstrated by its application in the Severe Asthma Research Program (SARP), [98]. The k-means algorithm is an iterative algorithm that aims to partition n objects into k unique clusters, $C_1, ..., C_k$, where $(k \leq n)$. Each object is assigned to the cluster with the nearest centroid such that the within-cluster *sum of squared error* is minimised. The difference between an object $\mathbf{p} \in C_i$ and the representative centroid $\mathbf{c}_i$ is represented by the square Euclidean distance $dist_{Eucl.}(\mathbf{p}, \mathbf{c}_i)$ between them. The aim is to minimise the objective function defined by Hartigan and Wong [99],

$$E = \sum_{i=1}^{k} \sum_{\mathbf{p} \in C_i} dist_{Eucl.}(\mathbf{p}, \mathbf{c}_i), \tag{3.8}$$

where E represents the sum of absolute error for all objects $\mathbf{p}$ and $\mathbf{c}_i$ is the representative centroid of cluster C_i.

K-Medoids and partition around medoids (PAM)

Although the k-means algorithm is regarded as a popular approach, the presence of outliers can alter the mean value of a cluster and as a result, the cluster centroid will be pushed closer to the outlier. The k-medoids approach addresses this shortcoming by selecting a more representative medoid as a reference point instead of calculating the arithmetic mean of the objects in a cluster. The Partition around medoids (PAM) algorithm, an implementation/realisation of the k-medoids algorithm, is slightly modified from the k-means algorithm, [99]. While popular, the k-means algorithm uses the Euclidean distance, which cannot be calculated with categorical data and is therefore not considered a possible approach in this chapter, [99, 100]. The PAM algorithm is a widely applied approach for mixed-variable data and is less sensitive to outliers than algorithms that use means, since it chooses medoids as cluster representatives, [100]. PAM differs from k-medoids in its methodology and not the choice of distance measure chosen.

Given dataset D, the PAM algorithm uses an iterative approach to partition n objects into k unique clusters, $C_1, ..., C_k$. The difference between an object $\mathbf{p} \in C_i$ and the representative medoid $\mathbf{m}_i$ is represented by $dist(\mathbf{p}, \mathbf{m}_i)$. This distance is usually calculated using one of the measures introduced earlier, depending on the type of data being clustered. For example, Euclidean distance is often used for continuous data, while Wishart distance is commonly applied to mixed variable data. The medoids of each cluster are calculated by minimising an objective function defined by Kaufman and Rousseeuw (1990), [101].

$$E_{PAM} = \sum_{i=1}^{k} \sum_{\mathbf{p} \in C_i} dist(\mathbf{p}, \mathbf{m}_i), \tag{3.9}$$

where E_{PAM} represents the sum of absolute error for all objects $\mathbf{p}$ and $\mathbf{m}_i$ is the representative medoid of cluster C_i.

The PAM algorithm is a flexible approach given that it can be implemented with any appropriate distance matrix while also robust to outliers in the data. However, the PAM algorithm suffers from a complexity problem (high run time cost) and does not scale well for large datasets, where the time complexity is in order of $O(k(n-k)^2)$. To address the issue of time complexity, Clustering LARge Applications (CLARA)[102] and Clustering Large Applications based on RANdomized Search (CLARANS)[103] have been proposed. Where CLARA provides an improvement over the PAM algorithm by locating medoids in only a sample from the data. The underlying idea of CLARA

states that if subsamples are sufficiently random, then the medoids of the subsample should be a strong approximation of the medoids in the complete dataset. The size of each subsample is suggested to be $(40 + 2k)$ objects, where k represents the target number of clusters, [104].

Cluster validation indices play a crucial role in evaluating the quality and validity of the clusters formed. The average silhouette width (ASW)[105] has previously been used and recommended, in the context of wheezing phenotypes, as a cluster validation index to estimate the number of clusters with PAM, [47, 101]. The ASW does not depend on statistical model assumptions.

The ASW is defined as follows. Let i be an object which belongs to cluster A and let C represent a cluster that does not include object i. Further, $a(i)$ represents the average dissimilarity between i and all other objects in A. Let $d(i, C)$ be the average dissimilarity between i and all objects in C. Upon calculating $d(i, C)$ for all clusters $C \neq A$, the smallest of those values is selected and denoted by:

$$b(i) = \min_{C \neq A} d(i, C) \ . \tag{3.10}$$

Note that $b(i)$ depends on the presence of clusters other than A. Therefore, it is essential to maintain the assumption that the number of clusters exceeds one, [105]. The silhouette width, $s(i)$, is defined in as

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \tag{3.11}$$

where $s(i)$ ranges from -1 to 1. A higher value within the range -1 to 1 suggests a strong alignment of the object with its designated cluster than to objects of the closest other cluster. In the case of a cluster that contains only a single object, $s(i)$ is assigned the arbitrary value of 0, [105].

While statistical judgements inform the optimal number of classes, reliance should not solely be on the ASW, but also visualisations of the internal structure to check for within-class homogeneity, intra-class separation, and guidance from literature on previously derived wheeze clusters. Clinical judgement is an integral part of the phenotype derivation process.

At this point, it is relevant to refer back to the longitudinal nature of the wheeze data which spans across 14 visits to track individual episodes. Applying conventional partitioning algorithms to obtain a clustering solution, such as k-means, on the DCHS wheeze data poses a non-trivial challenge. This is because partitioning algorithms require a similarity metric to compare pairs of binary wheeze profiles (temporal sequences) from subjects. The clustering algorithms above would treat the repeated measurements from subjects at the 14 different visits as 14 independent variables. Two measures that can be used are cross-correlation and dynamic time warping (DTW), [106]. Cross-correlation is used to measure the similarity between two sequences as a function of the lag of one relative to the other. DTW is an existing algorithm, applied in time series analysis, designed to quantify the similarity between two temporal sequences, [107]. However, it has been

noted that dynamic time warping does not apply itself well given a sparse series, [106]. Instead of turning directly to the widely applied model-based clustering solutions [16] to identify clusters of wheezing, attention is directed towards clustering summary indicators of longitudinal wheeze profiles using partitioning methods in Section 3.5.

3.5 Extending the PAM clustering approach

In this section, a simple, but effective approach is discussed to identify homogenous classes based on longitudinal binary indicators introduced by Haider et al (2022), [47]. The traditional PAM approach can be extended by adding an additional step to the distance-based clustering procedure between the primary data and the distance matrix in which longitudinal data points are summarised using a set of multi-dimensional indicators as illustrated further down in Figure 3.3. The summarised set of multi-dimensional indicators is used as input features for clustering. By doing so, this aims to create a more precise classification of the longitudinal data into distinct homogeneous classes. This extended PAM approach not only retains the simplicity of the original method but also leverages the added step to handle the complexities inherent in longitudinal binary data more effectively.

3.5.1 Multi-dimensional indicators

By extending the PAM approach, the aim is to summarise longitudinal observations of the presence/absence of a particular outcome of interest using multi-dimensional indicators for each child which are calculated using the primary data. In a clinical setting, the primary data would represent binary responses captured at child visits over an observational period. Describing the multi-dimensional indicators adequately is an important step since these indicators will later be used as input data into a distance matrix before applying the PAM algorithm. It is later shown in Table 3.7.1 how different subsets of indicators would produce different optimal clustering solutions.

To summarise longitudinal data of a clinical nature, most indicators can generally be divided into two groups, (1) those that represent the prevalence of a clinical outcome and (2) those that represent the time component of a clinical outcome. Here, the time component could be represented by the initial point of progression or decline of the clinical outcome. The purpose of the indicators is to summarise the evolution of a disease in a multidimensional way. However, by summarising the data in this way the longitudinal nature observed in the data is removed along with the statistical properties which accompany this. In a clinical setting, it is reasonable to include at least one indicator which summarises the time component for an event of interest, for example, the age of the first event. By including a time component in addition to a prevalence component, the data is allowed to be separated on both prevalence and temporal scales. In the case of the binary clinical outcome of wheeze, it is recommended to code the multi-dimensional indicators to include the following indicators[47]:

- Age of the first wheeze episode;
- Age of the last wheeze episode;
- Total number of separate wheeze episodes over the observation period (5-years);
- The length of the longest wheeze spell based on the number of consecutive episodes of wheeze;
- The total number of separate wheeze spells;

- Episode characteristics, referred to as a spell type, contrasting no wheeze to single or multiple intermittent episodes (defined as 0=healthy (no spell), 1=single spell, 2=intermittent spells).

The interpretation of clustering results can be derived from analysing the distribution of multi-dimensional indicators across the various clusters. For instance, if the total number of separate wheeze episodes during the observation period is significantly higher in one cluster compared to the others, this cluster can be characterised as representing recurrent or severe wheezing. Consequently, careful consideration is necessary when determining the number of indicators to include in the analysis, as high-dimensional feature spaces can complicate the interpretation of clustering outcomes.

Table 3.1 shows how the 6 multi-dimensional variables used in the DCHS were derived from the binary wheeze variables for 5 subjects. A spell is defined as beginning when a wheezing episode is first observed and ending when the next non-wheezing event is observed. In the example shown in Table 3.1, spell lengths can range from one to 14 consecutive time points, and individuals can experience multiple spells over 5 years.

Table 3.1: Derivation of wheezing indicators

Time point	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Time of the first wheeze episode	Time of the last wheeze episode	Total number of separate wheeze episodes	Length of the longest wheeze spell	Total number of separate wheeze spells	Spell type
Subject 1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	1	3	3	3	1	Single
Subject 2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	No wheeze
Subject 3	0	0	0	0	0	0	0	1	1	0	0	0	0	0	8	9	2	2	1	Single
Subject 4	0	0	0	1	1	0	0	1	1	1	0	0	0	0	4	10	5	3	2	Intermittent
Subject 5	0	0	0	1	0	0	1	0	0	1	1	0	0	0	4	11	4	2	3	Intermittent

3.5.2 Procedure

This section describes in more detail the procedure used to extend the PAM approach to longitudinal data. Consider the format of the primary input data illustrated by Figure 3.3 (left), where each row labelled $s_1, ..., s_n$ represents a row vector unique to each subject comprised of p binary variables with value 1 (presence) or 0 (absence) measured at different study visits.

The first step under the extended PAM approach is completed by summarising each subject-specific row vector using a set of multidimensional indicators which are considered to be the most appropriate. During the second step, the distance between every pair of rows of the multidimensional indicators, Figure 3.3 (centre), is calculated and stored in a distance matrix. The distance measure that will be used will be determined by the data types of the multi-dimensional indicators. Lastly, the PAM algorithm, introduced in Equation 3.9, is used to identify homogenous classes within the data sample by mapping a distance matrix into a user-specified number of classes.

The approach introduced is described as a feature-based approach. This can be seen as a specific instance of distance-based clustering, utilising a distance function customised for the particular

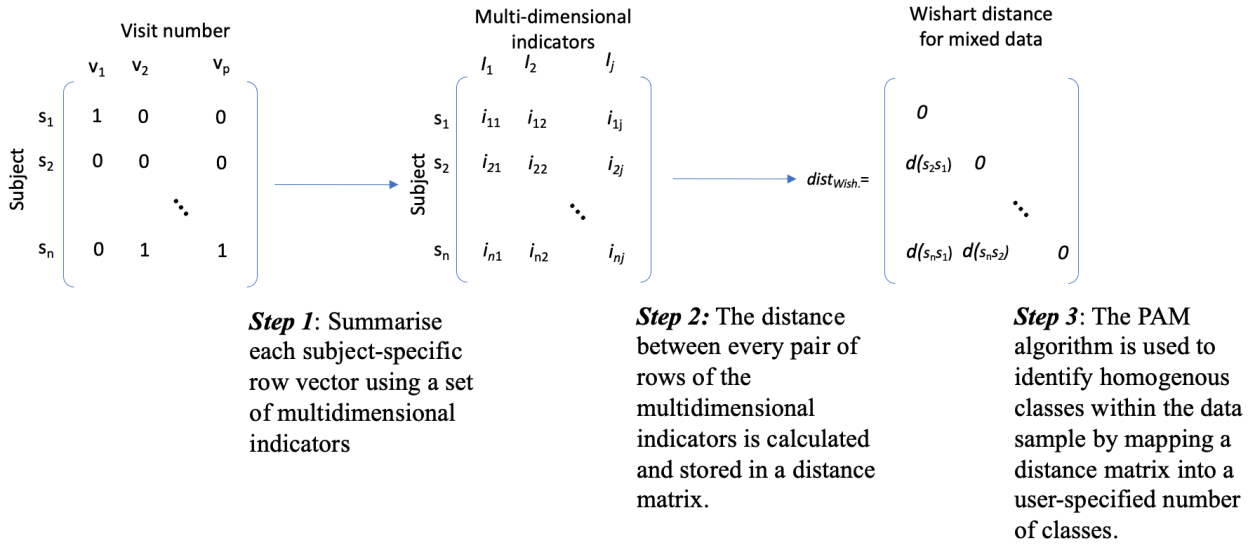


Figure 3.3: The format of the primary input data matrix (left), multidimensional indicators (centre), and the distance matrix of pairs (right) under the PAM approach.

domain of paediatrics research. A benefit of applying this approach is that researchers can easily integrate various features of clinical interest into the analysis. One drawback compared to model-based clustering approaches such as LCA which makes use of maximum likelihood estimation is that the reliability of the trajectories depends on the amount of data available per trajectory. Therefore, a larger number of observations per subject is necessary to achieve more accurate results.

3.6 Model based clustering

Many common clustering methods, such as the PAM algorithm, are based on distance or dissimilarity between data points. Another clustering approach commonly employed is model-based clustering, which is frequently utilised in the identification of wheeze and asthma phenotypes, [16]. Under a model-based clustering approach, it is assumed that there is a mixture of underlying probability distributions which generate the data.

LCA was introduced by Lazarsfeld (1995) as an approach to cluster individuals based on dichotomous observed variables, [108]. LCA yields a probabilistic clustering approach, as opposed to clustering approaches such as k-means and PAM, which are distance-based techniques. The practical applications of latent class models were made more significant by Goodman (1974) who provided the basis for obtaining maximum likelihood estimates of the model parameters, [109]. Bartholomew & Knott (1999) provided a framework which classifies latent variable models into four different classes of models based on the manifest and latent variables, as introduced by Table 3.2, [110]. Croudace (2003) employed standard latent class analysis to classify subjects based on longitudinal binary data, a methodology that has since been widely adopted by numerous researchers, [111]. In latent variable models, a manifest variable (or indicator) is a variable which can be directly measured or observed. Given its widespread application across various studies, the LCA approach can be regarded as the gold standard for identifying wheezing or asthma phenotypes, [16].

Table 3.2: Four classes of latent variable methods by Bartholomew and Knott (1999)

		Manifest Variables	
		Continuous	Categorical
Latent Variables	Continuous	Factor analysis	Latent trait analysis
	Categorical	Latent profile analysis	Latent class analysis

Given that both the latent and manifest variables are categorical while deriving wheezing phenotypes, the work in this chapter will solely focus on the theory and application of latent class analysis.

Given a set of 10 binary indicator variables, observing 2^{10} unique patterns of responses from subjects would be possible. Therefore, there could be a total of 2^{10} different clusters for subjects. However, the interest is in identifying a smaller number of C homogenous classes within the data. Latent class analysis (and other clustering approaches) allow us to do this by grouping the observed binary response patterns into C different classes, where $C < 2^{10}$. For example, the binary response patterns [0000001111] and [0000011110] could be assigned to the same class although the patterns are not completely identical, but rather similar. Each latent class is therefore characterised by *profiles* rather than exact binary matches between responses. For example, a latent class may be observed which contains subjects with a high prevalence of a disease or a latent class of healthy subjects with

a low prevalence of a disease.

Consider a vector of response patterns which is represented by $\mathbf{Y}$; y represents a particular response pattern. The probability of observing the response pattern y is defined by:

$$P(\mathbf{Y} = y) = \sum_{c=1}^{C} P(X = c) \times P(Y = y|X = c), \tag{3.12}$$

where C represents the number of latent classes from $c = 1, ..., C$ and X represents the categorical latent variable. Further, $P(X = c)$ denotes the probability of belonging to latent class c and $P(\mathbf{Y} = y|X = c)$ denotes the conditional probability of observing a particular response pattern y, given that the categorical latent variable X belongs to latent class c.

Latent class analysis requires the assumption of *local independence*, which assumes that observed variables within each latent class are independent of each other. Local independence assumes that any association between observed variables is explained by the latent class membership and not by direct relationships between the observed variables within the same class. The K manifest variables $y_1, ..., y_k$ are assumed to be mutually independent within each latent class which can be expressed as follows:

$$P(\mathbf{Y} = y|X = c) = \prod_{k=1}^{K} P(Y_k = y_k|X = c) \tag{3.13}$$

Combining the two equations 3.12 and 3.13 produces the following model for $P(\mathbf{Y} = y)$:

$$P(\mathbf{Y} = y) = \sum_{c=1}^{C} P(X = c) \times \prod_{k=1}^{K} P(Y_k = y_k|X = c) \tag{3.14}$$

The parameters of the model in Equation 3.14 can be estimated by using the maximum likelihood method. The general goal of latent class analysis is to assign subjects to a specific latent class. The probability of belonging to latent class c can be obtained using Bayes theorem:

$$P(X = c|\mathbf{Y} = y) = \frac{P(\mathbf{Y} = y|X = c) \times P(X = c)}{P(\mathbf{Y} = y)} \tag{3.15}$$

The most prevalent rule for allocating subjects to classes is known as the *modal* assignment rule, [112], where each subject is assigned to the latent class with the largest posterior membership probability $P(X = c|\mathbf{Y} = y)$ from equation 3.15. However, latent class assignment can also be done using *proportional* assignment rules, [113]. Proportional assignment rules use posterior probabilities from Equation 3.15 to assign subjects probabilistically, meaning they are not fully allocated to one class but rather weighted across all classes.

Model-based clustering techniques share many advantages over traditional distance-based techniques. Assignment to clusters is based upon model estimated membership probabilities and determining the number of clusters becomes a question of model selection, [114]. Further, model-based clustering allows the addition of covariates for cluster description and can accommodate non-Gaussian data. For advantages such as these, LCA is considered the gold standard in the clinical setting to identify wheeze and asthma phenotypes and has been widely applied, [16].

A variety of methods exist to select the optimal number of classes, while literature shows that the Bayesian Information Criterion (BIC) is the most popular method among researchers investigating wheeze and asthma phenotypes, [16]. The BIC [115] is given in Equation 3.16.

$$\text{BIC} = -2\ln(\mathcal{L}) + k\log(n) \tag{3.16}$$

Where, $\mathcal{L}$ is the maximised value of the likelihood function for the estimated model, n is the number of data points in the observed data, and k is the number of free parameters that need to be estimated. The BIC is computed for each model and one then selects the model with the lowest BIC value. Sophisticated models are subject to penalties to prevent over-fitting, while parsimonious models are rewarded, [115].

The Akaike Information Criterion (AIC), which is also used in the succeeding lung function chapter, is another popular method, [116, 117, 118]. The AIC is given in Equation 3.17

$$\text{AIC} = -2\ln(\mathcal{L}) + 2k \tag{3.17}$$

The AIC penalty term is smaller than in the BIC, where the AIC employs a complexity penalty of $2k$, whereas with BIC the penalty is $k\log(n)$. Parametric methods that have been used include bootstrapped likelihood ratio tests [39, 118] and likelihood ratio chi-squared test statistic, [119]. Alternative techniques for model selection include model entropy [39, 117, 118] and a dissimilarity index, [119].

LCA classifies subjects based on their observed data, typically binary outcome data. It assumes that within each population, diseased or non-diseased, the observed variables are independent. The problem of conditional dependence, where outcomes are not independent within a latent class, can lead to the identification of spurious latent classes that do not exist at the underlying level, [120]. To address the problem of heterogeneity within classes, a solution was proposed by Qu et al. (1996), [121]. Their approach involves combining a latent class model with a random effect to account for the heterogeneity. The probabilities associated with the model are transformed to the probit or logit scale, and a normally distributed random effect is introduced for each subject. Afterwards, the model is transformed back to probabilities.

Random LCA expands on the traditional LCA by incorporating random effects, which capture variability that cannot be explained by the latent classes alone. In this model, the introduction of subject-level random effects allows outcomes to vary between individuals within the same latent class, thus relaxing the assumption of local independence.

The random effect helps capture the variation or differences within a latent class, allowing for a more flexible and realistic representation of the underlying structure. The difference between LCA and random LCA therefore lies in their treatment of the assumption of conditional independence and model flexibility.

The probability of observing the outcomes y_{ij} (the j^{th} binary outcome for subject i) given the latent class c and random effect λ is written as

$$\mathrm{P}\left(y_{i1}, y_{i2}, \ldots, y_{ik} \mid c_i = c, \lambda_i = \lambda\right) = \prod_{j=1}^{k} \pi_{icj}^{y_{ij}} \left(1 - \pi_{icj}\right)^{1-y_{ij}} \tag{3.18}$$

The probability π_{icj}, which represents the likelihood of the j^{th} outcome for subject i in latent class c, can be modeled using either a logit or probit link function:

For the logit scale:

$$\pi_{icj} = \frac{\exp(\alpha_{cj} + \beta_{cj}\lambda_i)}{1 + \exp(\alpha_{cj} + \beta_{cj}\lambda_i)} \tag{3.19}$$

For the probit scale:

$$\pi_{icj} = \Phi(\alpha_{cj} + \beta_{cj}\lambda_i) \tag{3.20}$$

where α_{cj} is the baseline logit or probit for class c on outcome j, and β_{cj} is the loading for the random effect λ_i, which follows a normal distribution $\lambda_i \sim N(0, \sigma^2)$.

The joint likelihood of observing all outcomes for subject i can then be expressed by integrating over the random effects and summing over the latent classes, as shown earlier in the general likelihood structure in Equation 3.18.

3.7 Results: Wheezing phenotypes in early childhood

From this point onward, this chapter will concentrate on the application, results, and discussions related to identifying wheezing phenotypes and their associated risk factors. To uncover distinct patterns of wheezing in children, clustering was performed on 950 (83%) of 1143 children from the DCHS who had complete wheezing data collected across 14 time points. The objective was to identify homogeneous wheezing phenotypes. Excluded children had similar characteristics to those included in the analysis, with the exception of higher socioeconomic status and a lower prevalence of antenatal smoking in the excluded group, Appendix A.1. Several algorithms for identifying clusters were reviewed, (1) partitioning around mediods (PAM) using multi-dimensional indicators, (2) latent class analysis (LCA), and (3) random LCA. The results will be introduced and discussed in this section.

3.7.1 Partitioning around mediods (PAM) and multi-dimensional indicators

Using a PAM approach tailored for longitudinal data, observations of wheeze presence or absence were summarised using 6 indicators, emanated from analyses by the UK STELAR consortium, Haider et al. (2022), [47]. These indicators include: (1) age at the first recorded episode (2), age at the last recorded episode; (3) total number of separate records over the observation period; (4) duration of the longest spell based on the number of consecutive records of wheeze; (5) total number of spells; (6) spell type, defined as 0 for no wheeze, 1 for a single spell, 2 for intermittent spells.

Based on the indicators listed above, the PAM algorithm was used to cluster the children into homogenous wheezing classes, referred to as phenotypes, by mapping a distance matrix into a user-specified number of wheezing classes. The Wishart distance matrix was used as input for the PAM algorithm applied to mixed variable data. The algorithm was run for 2-7 clusters, and the optimal number of clusters was chosen using the average of the silhouette width criterion. PAM models were fit in R version 3.6.3 (2020-02-29) using the *cluster* library, [122].

Findings related to the identified wheezing phenotypes

The structure of wheeze phenotypes was inferred in an unsupervised manner using the PAM algorithm, with the optimal solution suggesting 4 clusters based on the silhouette criterion, Table 3.3. The average silhouette width increases from 0.630 for two phenotypes to a peak of 0.710 for four phenotypes, before slightly decreasing afterwards. The derived wheezing phenotypes and their accompanying heat maps for the first 4 solutions are displayed in Figure 3.5 which shows the progression of moving from a single class to a 4-class solution.

Figure 3.5 (right) displays heatmaps representing the wheezing data, providing insights into the evolution of a positive wheezing diagnosis over time. The wheezing history of each child over 5 years will be plotted along the x-axis. The colour designates positive or negative diagnoses of wheezing and the rows are sorted by the number of wheezing episodes per child, as shown in Figure

Table 3.3: Average silhouette width for various phenotype sizes identified by the PAM model in Figure 3.5.

Number of Phenotypes	Average Silhouette Width
2	0.630
3	0.671
4	0.710
5	0.674
6	0.701
7	0.702

3.5. The darker colours of the bars or lines represent the presence of wheeze and lighter colours represent the absence of wheeze. This colour-coding technique is employed consistently throughout this chapter and an example of this is illustrated in Figure 3.4.

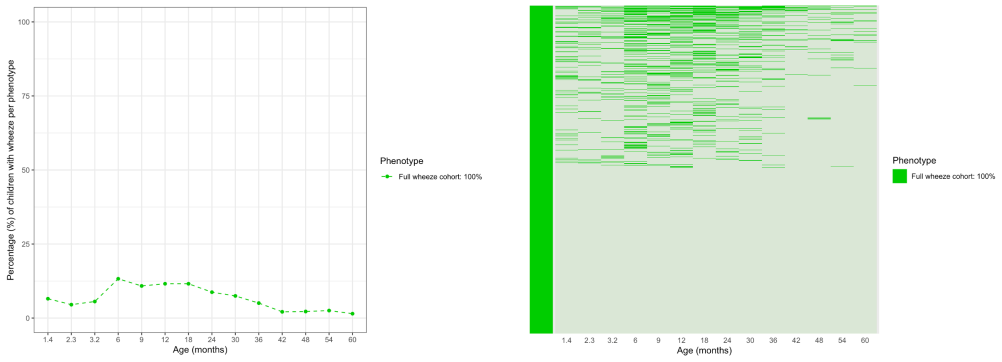


Figure 3.4: Colour coding used in the heatmaps.

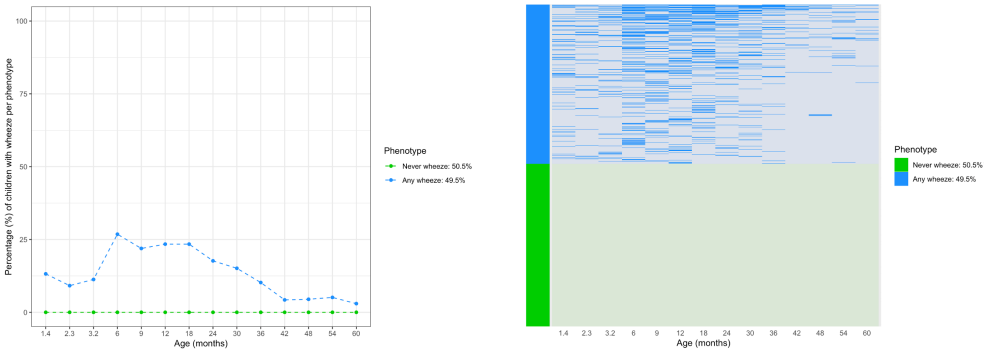
Disease prevalence plots illustrate wheezing profiles for each phenotype. Each wheezing profile shows the proportion of children with reported wheezing within each phenotype at different visits between birth and 5 years, as shown in Figure 3.5 (left). This is also referred to as outcome probabilities.

Figure 3.5 illustrates how the algorithm initially represents a dichotomous solution through a wheeze-positive and a wheeze-negative class. The wheeze-negative class remains present in the clustering solution, while the 3-class solution further subdivides the wheeze-positive class into a class of lower wheeze prevalence and another of higher wheeze prevalence. Up to the 3-class solution, the wheezing phenotypes have been defined by wheezing prevalence, but as progression moves to the 4-class solution the wheezing phenotypes begin to become defined by both prevalence and temporal variation. The PAM approach tailored for longitudinal data consistently and accurately distinguishes between children with a history of wheezing and those without for k greater or equal to 2.

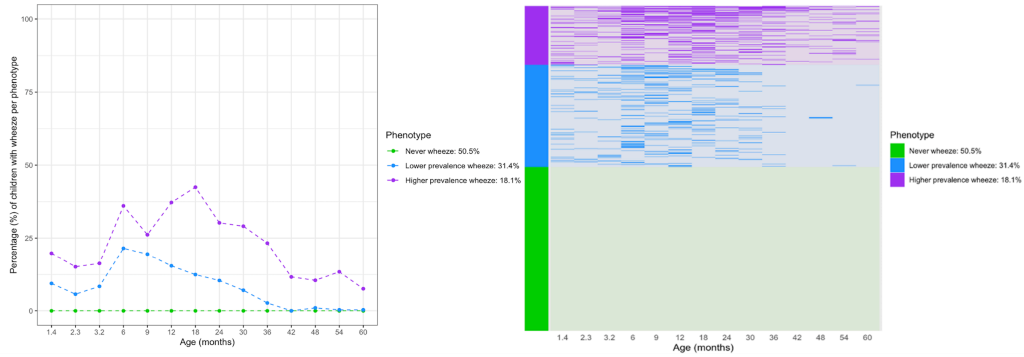
After examining the trajectories of wheeze prevalence, the age of onset, and the duration of wheeze prevalence for each cluster shown in Figure 3.5 d (left), the clusters (phenotypes) of the 4-class solution were characterised as 1) never wheeze (480/950, 50.5%); 2) early transient wheeze (215/950, 22.6%), in which wheeze was observed from 1.4 months, peaked at 6 months (29.8%) but remitted after 18 months; 3) late onset wheeze (104/950, 11.0%) with wheeze observed from 18 months, where it was the highest (34%), and declined to 4% by 5 years; and 4) recurrent wheeze (151/950, 15.9%) where wheeze was observed intermittently throughout. Intra-individual wheeze patterns



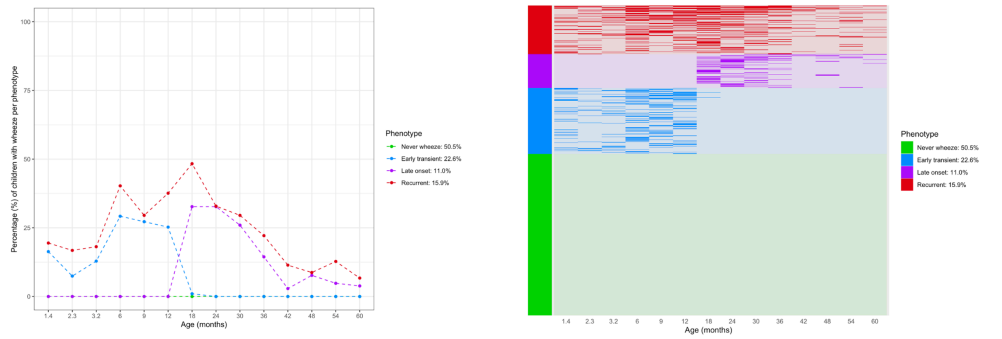
(a) Solution for k=1



(b) Solution for k=2



(c) Solution for k=3



(d) Solution for k=4

Figure 3.5: Percentage of children with reported wheeze in the first 5 years of life within each wheeze phenotype (left). Intra-phenotype individual wheeze patterns (right), stratified by phenotype; the darker colours of the bars or lines represent the presence of wheeze and lighter colours represent the absence of wheeze.

stratified by phenotype are visualised in Figure 3.5 d (right). Table 3.4 shows the distribution of the derived indicators stratified by phenotype. Recurrent wheeze was the only phenotype where children experienced >3 separate wheeze episodes, and spells lasting 3-8 consecutive time points. In contrast, early transient and late onset wheezing were characterised by less frequent, shorter, and a greater proportion of single spells.

(a) Total number of separate wheeze episodes

	Never wheeze		Early transient		Late onset		Recurrent	
0	480	(100.0%)	0	(0.0%)	0	(0.0%)	0	(0.0%)
1	0	(0.0%)	176	(81.9%)	80	(76.9%)	0	(0.0%)
2	0	(0.0%)	39	(18.1%)	22	(21.1%)	51	(33.8%)
3	0	(0.0%)	0	(0.0%)	2	(1.9%)	51	(33.8%)
4	0	(0.0%)	0	(0.0%)	0	(0.0%)	25	(16.6%)
5	0	(0.0%)	0	(0.0%)	0	(0.0%)	10	(6.6%)
6	0	(0.0%)	0	(0.0%)	0	(0.0%)	4	(2.6%)
7	0	(0.0%)	0	(0.0%)	0	(0.0%)	7	(4.6%)
8	0	(0.0%)	0	(0.0%)	0	(0.0%)	1	(0.7%)
9	0	(0.0%)	0	(0.0%)	0	(0.0%)	2	(1.3%)
Total	480	(100%)	215	(100%)	104	(100%)	151	(100%)

(b) Total number of separate wheeze spells

	Never wheeze		Early transient		Late onset		Recurrent	
0	480	(100.0%)	0	(0.0%)	0	(0.0%)	0	(0.0%)
1	0	(0.0%)	199	(92.6%)	94	(90.4%)	11	(7.3%)
2	0	(0.0%)	16	(7.4%)	9	(8.6%)	100	(66.2%)
3	0	(0.0%)	0	(0.0%)	1	(1.0%)	35	(23.2%)
4	0	(0.0%)	0	(0.0%)	0	(0.0%)	4	(2.6%)
5	0	(0.0%)	0	(0.0%)	0	(0.0%)	1	(0.7%)
Total	480	(100%)	215	(100%)	104	(100%)	151	(100%)

(c) Longest spell based on the number of consecutive episodes of wheeze

	Never wheeze		Early transient		Late onset		Recurrent	
0	480	(100.0%)	0	(0.0%)	0	(0.0%)	0	(0.0%)
1	0	(0.0%)	192	(89.3%)	89	(85.6%)	68	(45.0%)
2	0	(0.0%)	23	(10.7%)	15	(14.4%)	46	(30.6%)
3	0	(0.0%)	0	(0.0%)	0	(0.0%)	21	(13.9%)
4	0	(0.0%)	0	(0.0%)	0	(0.0%)	7	(4.6%)
5	0	(0.0%)	0	(0.0%)	0	(0.0%)	4	(2.6%)
6	0	(0.0%)	0	(0.0%)	0	(0.0%)	4	(2.6%)
8	0	(0.0%)	0	(0.0%)	0	(0.0%)	1	(0.7%)
Total	480	(100%)	215	(100%)	104	(100%)	151	(100%)

(d) Spell type

	Never wheeze		Early transient		Late onset		Recurrent	
Intermittent1	0	(0.0%)	16	(7.4%)	10	(9.6%)	141	(93.4%)
Single	0	(0.0%)	199	(92.6%)	94	(90.4%)	10	(6.6%)
No wheeze	480	(100.0%)	0	(0.0%)	0	(0.0%)	0	(0.0%)
Total	480	(100%)	215	(100%)	104	(100%)	151	(100%)

[1]intermittent is defined as at least 2 non-consecutive spells of wheeze of any length

Table 3.4: Distribution of the derived indicators stratified by phenotype (PAM 4-Class model).

3.7.2 Sensitivity analysis of the 4-class phenotype solution under PAM

This section investigated the sensitivity of the final clustering solution identified and described by the 4-class solution using the PAM approach by comparing the results under different conditions, specifically sample size and input indicators.

Several decisions have to be made to perform cluster analysis, such as the choice of the clustering method (e.g., PAM or LCA), the number of clusters, the dissimilarity function, and its parameters. A natural question that follows the identification of clusters is determining how stable the clustering solution is concerning changes in the input data. To address this, the stability of the clustering results is explored by examining how changes in the input indicators affect the final phenotypes identified by the PAM approach.

Within the PAM framework, investigation into the sensitivity of clustering results is initiated based on different subsets of the multidimensional indicators using Algorithm 1. The algorithm starts by selecting the full set of multi-dimensional indicators for all subjects, and then secondly selects one of the six indicators listed in Section 3.5.1 to be excluded, thereby creating a subset of indicators. The PAM algorithm was then used to create clusterings by mapping an appropriate distance matrix (see notes under Table 3.5) into 2-6 classes. This process was repeated 6 times, each time excluding a different indicator. The ASW for each subset of indicators is shown in Table 3.5.

Algorithm 1 Stability of the PAM clustering results based on different subsets of multi-dimensional indicators.

procedure
  N ← Full set of subjects included in the analysis
  for `sample.indicator` *in* $[ind.1, ind.2, ind.3, ind.4, ind.5, ind.6]$ **do**
    data.sample ← Select all columns in the data matrix minus `sample.indicator`
    for `number.phenotypes` *in* $[2:6]$ **do**
      PAM model with classes equal to `number.phenotypes` and data equal to `data.sample`
      Store the silhouette width from the model for each `number.phenotypes`
    end for
  end for
  return The average silhouette width and standard deviation for each phenotype size within `sample.size` across all iterations.
end procedure

Table 3.5 demonstrates that the exclusion of indicators 4 (the length of the longest wheeze spell), 5 (the total number of separate wheeze spells), or 6 (episode characteristics contrasting no wheeze to single or multiple intermittent episodes) changes the optimal number of phenotypes away from 4. These findings show that wheezing spells, an indicator of severity, represent an influential input characteristic within the DCHS cohort. The average silhouette width values suggest indicators play an important role in defining distinct wheezing phenotypes. This underscores the importance

Table 3.5: Wheezing classes identified by the PAM model - Average silhouette width for indicator subsets

Number of Phenotypes	Excluding *Indicator 1*	Excluding *Indicator 2*	Excluding *Indicator 3*	Excluding *Indicator 4*	Excluding *Indicator 5*	Excluding *Indicator 6*
2	0.634	0.616	0.603	0.664	0.656	0.687
3	0.664	0.676	0.651	0.723	0.699	0.672
4	0.711	0.729	0.705	0.727	0.728	0.675
5	0.679	0.704	0.668	0.742	0.746	0.696
6	0.699	0.721	0.697	0.689	0.756	0.705

The Euclidean distance matrix was used as input to the PAM algorithm for numeric data while excluding the categorical indicator (6).
The Wishart distance matrix was used as input to the PAM algorithm for mixed variable data.

of including comprehensive wheezing-related data in clustering analyses to accurately capture the heterogeneity.

Although phenotypes are identified through a data-driven process, prior decisions need to be made such as the type and number of indicators to be included in the PAM algorithm which introduces a certain degree of subjectivity. Sensitivity analysis, using different subsets of indicators, showed that wheezing phenotypes are dependent on the indicators used as input into the PAM model, Table 3.5. Therefore, results obtained from different cohorts can only be comparable if the same set of indicators are used.

Results such as these are consistent with previous studies which illustrated that variations in input parameters, even within the same cohort, can lead to different results. An illustrative example can be found in two publications from the Severe Asthma Research Program (SARP), [123, 124]. Both papers made use of Ward's minimum-variance hierarchical clustering method by using an agglomerative (bottom-up) approach and Ward linkage and reported different clustering results. The difference between the paper's approaches lies in the data pre-processing methods and the cluster input parameters.

Prosperi et al. (2013) proposed that the differences in resulting clusters across various studies can be attributed to the diverse choices made by investigators regarding factors, the encoding, categorisation, and transformation of variables, as well as the methodologies employed, [125]. A useful approach to evaluate the robustness of the applied methods is through bootstrapping and multiple repetitions, although this does not inherently lead to more credible overall outcomes.

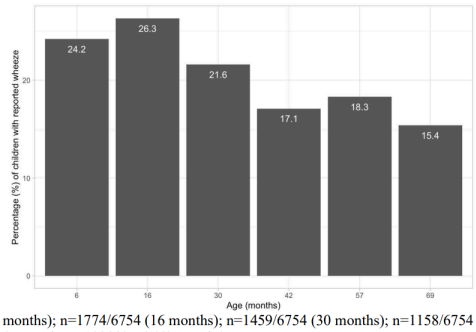
To summarise, not all indicators contribute equally to the clustering process, and irrelevant indicators can distort the output. Indicators should therefore be carefully selected.

3.7.3 DCHS versus Avon Longitudinal Study of Parents and Children (ALSPAC)

The 4-class phenotype structure under the PAM approach was compared with the same analysis in the Avon Longitudinal Study of Parents and Children (ALSPAC), [36]. ALSPAC is a UK birth cohort study established in 1991 in Avon, UK, [36]. It recruited pregnant women with expected delivery dates between April 1, 1991, and December 31, 1992.

ALSPAC was used as a comparator cohort since this is one of the largest UK cohorts with the greatest number of data collection time points in the first 5 years and therefore provides a robust comparator in a HIC. Identical methods were used to derive wheeze phenotypes in ALSPAC. A comparative analysis was conducted to compare the 4-class solution from DCHS with the 5-class solution from ALSPAC.

For ALSPAC, validated questionnaires were completed repeatedly from infancy through adolescence, [49]. Wheeze clusters were identified for a cohort of 6,754 children in ALSPAC, who possessed complete data regarding wheezing episodes at ages 6, 16, 30, 42, 57, and 69 months, [25]. This serves as a strong, though not perfect, comparison to the DCHS. Data from ALSPAC continues until 69 months of age with 6 visits, whereas the DCHS covers up to 60 months of age with 14 visits. The prevalence of children experiencing at least one wheezing episode in this group closely resembled the rate observed in the DCHS cohort, with 3,489 (52%) out of 6,754 children in ALSPAC and 470 (49%) out of 950 children in DCHS. Figure 3.6 displays the prevalence of current wheeze across 6 time points from ALSPAK, comparisons can be made to Figure 3.2 from DCHS.



n=1635/6754 (6 months); n=1774/6754 (16 months); n=1459/6754 (30 months); n=1158/6754 (42 months); n=1233/6754 (57); months n=1038/6754 (69 months)

Figure 3.6: Point prevalence of current wheeze in children in the ALSPAK by age. See McCready et al. (2023).

However, at 6 months of age, approximately twice as many children in the ALSPAC study experienced a wheezing episode compared to those in the DCHS study. Specifically, 1635 out of 6754 children in ALSPAC (24%) had wheezing, while only 126 out of 950 children in DCHS (13%) reported episodes. By the age of 5, wheeze prevalence declined to 18% (1233 of 6754 children) in ALSPAC compared with 1% (14 of 950 children) in DCHS, Figure 3.2 and 3.6. A different pattern

of wheezing was identified in ALSPAC, with a 5-cluster solution considered optimal, Figure A.1. The clusters are as follows: never wheeze (3265 children, 48% of the total 6754), early transient (1596 children, 24%), late onset (551 children, 8%), intermittent (783 children, 12%), and persistent (559 children, 8%), as illustrated in Figure 3.7. A cluster unique to the ALSPAC study, persistent wheeze, was defined by an extended period of wheezing across any time point. The other four clusters appeared to be similar; however, children in the early transient wheeze cluster in ALSPAC experienced a longer duration of wheezing before remission compared to those in the DCHS (42 months vs 18 months), and late-onset wheezing began later in ALSPAC than the DCHS (42 months vs 18 months).

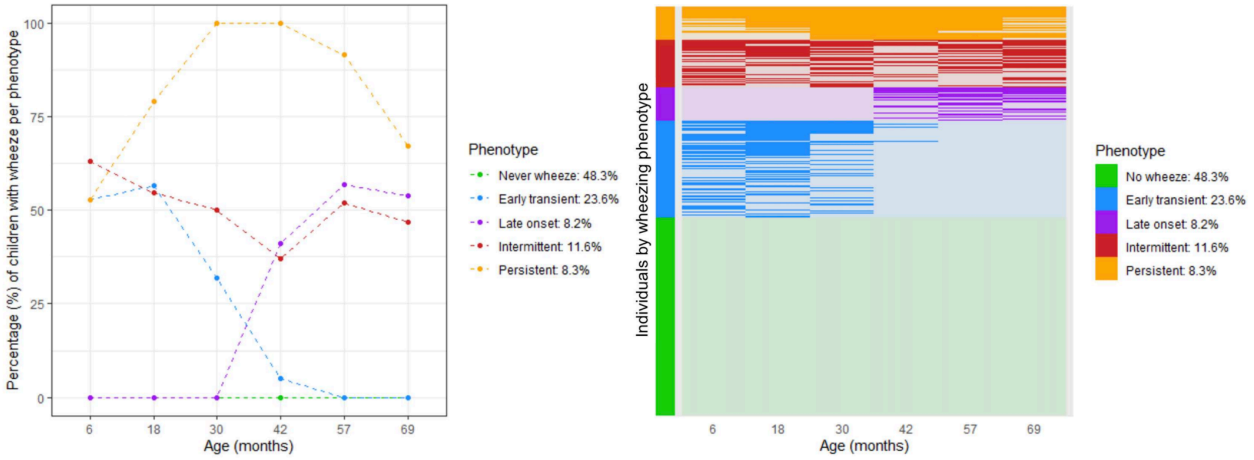


Figure 3.7: ALSPAC cohort - Percentage of children with reported wheeze within each wheeze phenotype (left). Intra-phenotype individual wheeze patterns (right), stratified by phenotype; the darker colours of the bars or lines represent the presence of wheeze and lighter colours represent the absence of wheeze. See McCready et al. (2023).

Wheeze frequency in the DCHS cohort was lower than the ALSPAC cohort in the UK and wheezing prevalence declined more rapidly. Four wheeze clusters were identified in DCHS, which displayed differences in the age of onset and absence of a persistent phenotype compared with those in ALSPAC. The wheezing phenotypes identified in this LMIC setting may appear similar to those derived from data-driven analyses in birth cohorts in high-income countries. However, most cohorts from high-income countries identified five wheeze phenotypes, [48, 47, 39]. Although the phenotypes seemed similar, these differed in DCHS and ALSPAC. For example, wheezing in the early phenotype resolved earlier in DCHS than in ALSPAC (18 months vs 42 months), late wheezing started earlier in DCHS (12 months vs 42 months), and no persistent wheeze phenotype was identified in DCHS. These differences indicate a quicker cessation of early wheezing, but a younger age for the development of late wheezing in DCHS than in ALSPAC Such differences might be partly due to the high burden of RSV-LRTI as a key precipitant of wheezing phenotypes in this setting, with the

highest prevalence in infants. Global estimates report a major proportion of respiratory syncytial virus-LRTI to occur during infancy, with the burden heavily concentrated in LMICs, [22].

Since ALSPAC identified 5 phenotypes, a comparison was also made to the 5-phenotype solution from the DCHS. The DCHS's 5-phenotype solution resulted in a less desirable ASW figure of 0.674, Table 3.3, compared to the suggested optimal 4-cluster solution, which had an ASW of 0.71 as previously compared to ALSPAC in the paragraphs above. The 5-class phenotype solution is shown in Figure 3.8. The additional phenotype not captured in the 4-class solution of the DCHS is represented by Class 2 in Figure 3.8. This phenotype characterises children who primarily reported wheezing between 9 and 18 months. Class 2 in Figure 3.8 does not exhibit any similarities to the phenotypes observed in the ALSPAC study.

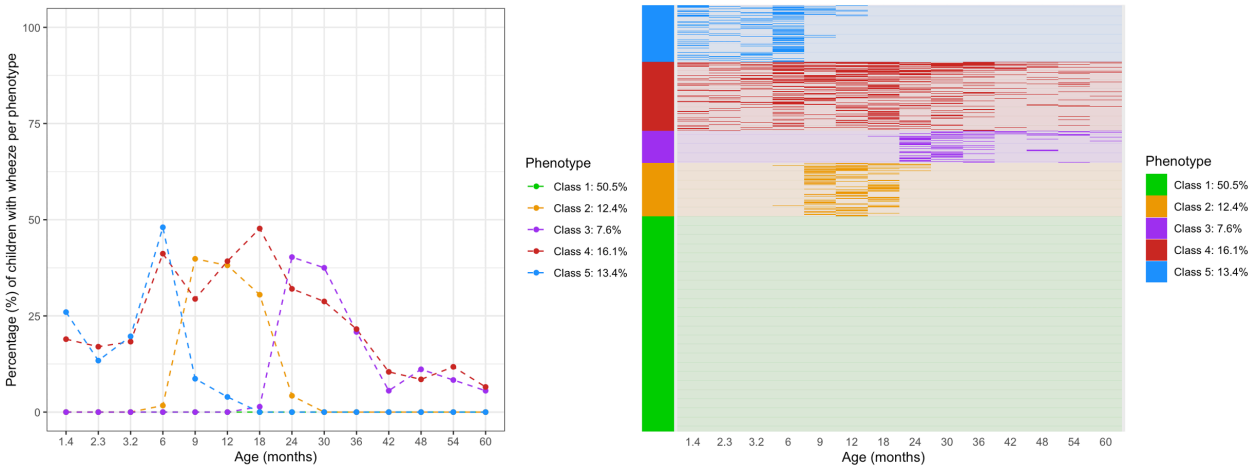


Figure 3.8: DCHS cohort 5-class solution - Percentage of children with reported wheeze within each wheeze phenotype (left). Intra-phenotype individual wheeze patterns (right), stratified by phenotype; the darker colours of the bars or lines represent the presence of wheeze and lighter colours represent the absence of wheeze.

3.7.4 Latent class analysis (LCA)

A comparison was made between the results from the PAM models and those from the widely applied LCA approach. LCA and PAM are both popular unsupervised approaches which are used to partition subjects into homogeneous groups based on dichotomous data that indicate the presence or absence of characteristics of interest. LCA fits a probabilistic model where it is assumed that a mixture of underlying probability distributions generated the data. LCA was performed to cluster the cohort into homogenous wheezing phenotypes, where models with varying numbers of latent classes (2-4) were compared. The optimal number of phenotypes was selected based on the lowest BIC value. LCA models were fit in R version 3.6.3 (2020-02-29) using the packages poLCA, [126]. A comparison was made between the LCA and PAM approaches in terms of the optimal number of wheezing phenotypes and the temporal differences in mean phenotype trajectories.

Using LCA, an optimal solution was obtained with 2 unique classes, shown in Figure 3.9, in contrast to the optimal solution of 4 classes obtained from the PAM model, shown previously in Figure 3.5. LCA appeared to separate wheeze phenotypes based on wheeze prevalence within a class, whereas PAM separates wheeze phenotypes based on both wheeze prevalence within the class and the age of the first or last onset of wheeze - a function of the chosen input characteristics. The first split made by the LCA approach does not clearly distinguish between a disease-positive and disease-negative class, where the LCA model assigned both wheezing and non-wheezing children to the same class. Both LCA (phenotype 2 from the 2-class model) and the extended PAM approach (phenotype 4 from the 4-class model) proved capable of identifying the most severe recurrent wheezing phenotype. Although these two methods proved capable of identifying wheeze phenotypes with varying degrees of success, the results presented here suggest that they are not interchangeable.

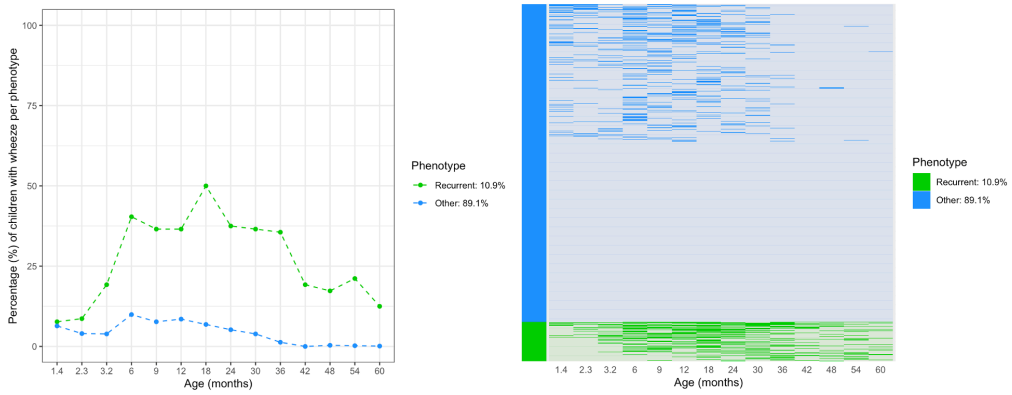
Table 3.6: BIC values for LCA and LCA RE models.

Number of Phenotypes	LCA	LCA RE (Common)	LCA RE (by class)
2	6048.8	5992.5	5997.4
3	6092.9	6065.3	6067.1
4	6164.8	6131.9	6144.7

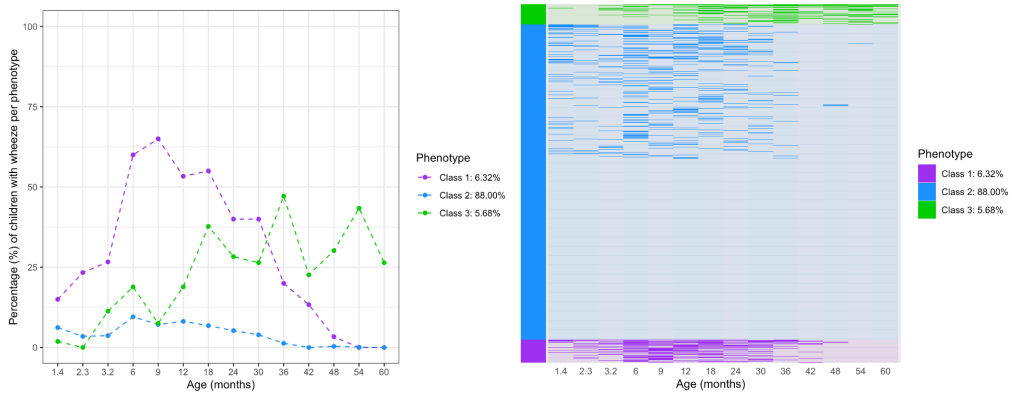
RE: Random effect; 'by class': random effect loading(s) vary by class.

The analysis was further extended to investigate output from random effects latent class models. Random effects models provide a better fit than standard latent class models, as shown in Table 3.6. Allowing the loadings to vary between classes provided no improvement in model fit. The BIC index increased with each additional class, implying that no optimal solution was evident, Table 3.6. The lowest BIC value was observed for 2 phenotypes. The profiles of wheezing patterns derived from the best performing 2-class LCA random effect model are presented in Figure 3.10.

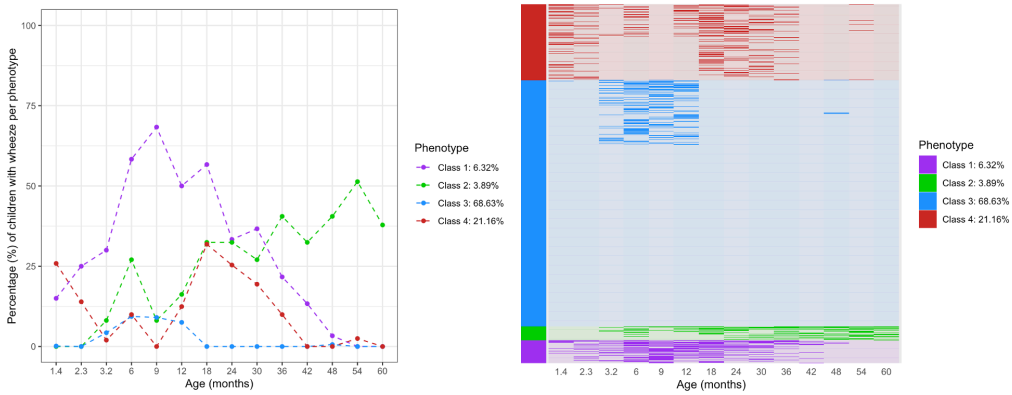
While comparing the classes between LCA and random LCA (Common), it was found that the two different methods did not show good agreement, as shown in Table 3.7. The row percentages from



(a) Solution for k=2



(b) Solution for k=3



(c) Solution for k=4

Figure 3.9: Identified by the LCA model. Percentage of children with reported wheeze in the first 5 years of life within each wheeze phenotype (left). Intra-phenotype individual wheeze patterns (right), stratied by phenotype; the darker colours of the bars or lines represent the presence of wheeze and lighter colours represent the absence of wheeze.

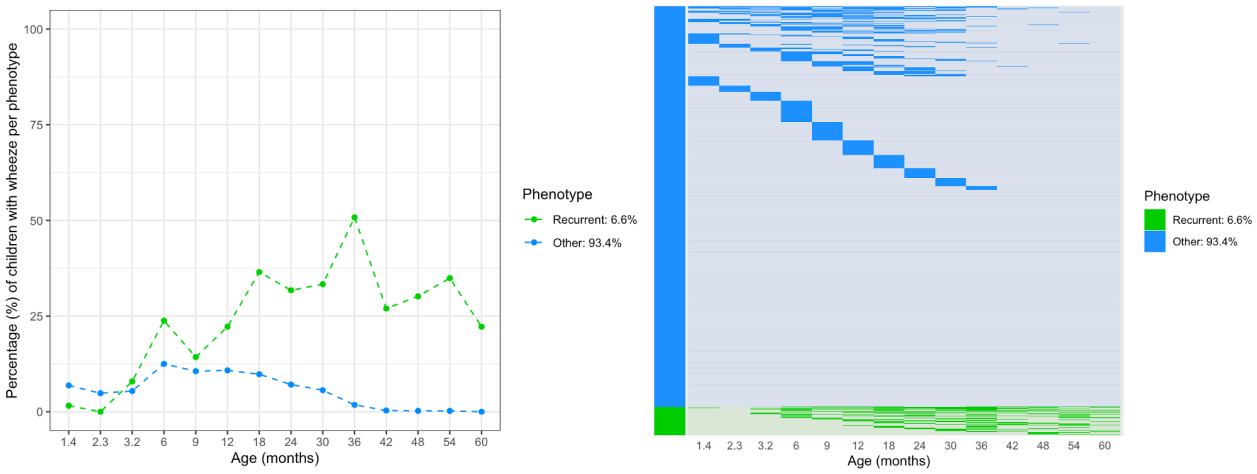


Figure 3.10: Percentage of children with reported wheeze in the first 5 years of life within each wheeze phenotype (left). Intra-phenotype individual wheeze patterns (right), stratied by phenotype; the darker colours of the bars or lines represent the presence of wheeze and lighter colours represent the absence of wheeze.

Table 3.7: A 2 by 2 cross-table of class assignments between LCA and random LCA (Common) models.

	LCA RE class 1	**LCA RE class 2**
LCA class 1	8 (7.69%)	96 (92.31%)
LCA class 2	55 (6.50%)	791 (93.50%)

Table 3.7 indicate that the majority of LCA class 1 is classified into LCA RE class 2, and similarly, most of LCA class 2 is classified into LCA RE class 2. The shape of the wheezing profiles differed between LCA and random LCA, as shown in Figures 3.9 and 3.10. The recurrent (green) profiles are different when comparing the 2-class LCA model to the 2-class random LCA model. Under LCA, the outcome probabilities showed a downward trend from 6/12 months to 60 months, while under random LCA, they showed an upward trend. The random effect component appears to shift the prevalence to peak later. It shows the accumulation of repeated wheeze episodes within subject better.

Limitations of LCA within the DCHS

In a simple scenario, a clustering solution should be able to divide the subjects into two classes: healthy subjects and unhealthy subjects. However, upon applying LCA to the DCHS data, the model was unable to distinguish between healthy subjects and those who had reported wheeze.

Unlike the PAM approach, which would perfectly separate the healthy from the wheezing classes, LCA techniques identified classes of high and low prevalence due to its probabilistic approach.

The analysis conducted in this study suggests that the inability of LCA to distinguish between healthy and unhealthy subjects could be partly due to infrequent or singularly occurring patterns of wheezing. This is believed to be a consequence of having data recorded at 14 different time points, which leads to the possibility of observing 2^{14} different response patterns. This represents a large number of possible patterns, and many of them are likely very rare. A total of 151 distinct patterns of wheeze were observed in the DCHS, the top fifteen patterns over 14 time points are shown in Table 3.8. Additionally, it was noted that random LCA struggles to differentiate between healthy and unhealthy subjects when the wheezing prevalence was low (for example, one wheezing observation over 14 time points), even if the pattern of wheezing appears frequently. Table 3.8 shows that a high prevalence of a single episode was observed at different time points and this is reflected by the blue profile in Figures 3.9 and 3.10.

Table 3.8: Top fifteen wheeze patterns in the DCHS

Pattern	Frequency	cumulative %
00000000000000	480	50.52%
00010000000000	47	55.47%
00001000000000	40	59.68%
00000100000000	33	63.15%
00000010000000	29	66.21%
00000001000000	22	68.52%
10000000000000	21	70.73%
00100000000000	20	72.84%
00000000100000	18	74.73%
01000000000000	13	76.11%
00001100000000	9	77.05%
00000000010000	8	77.89%
00000101000000	7	78.63%
00011000000000	7	79.36%
00000011000000	6	80.00%
.	.	.

A unique pattern was observed in 106 subjects.

These points were illustrated by testing a simple 2-class random LCA solution on two different subsets of data from the DCHS. In the first example, it is shown how random LCA struggles with low prevalence by testing a 2-class model on three frequently occurring patterns from the DCHS: a never wheezing pattern (00000000000000) with 480 occurrences, a single wheezing pattern (00010000000000) with 47 occurrences, and a single wheezing pattern (00001000000000) with 40

occurrences. Figure 3.11 shows the profiles from the model output in this scenario, demonstrating that random LCA does not perfectly separate healthy and wheezing subjects when the prevalence of wheezing is low.

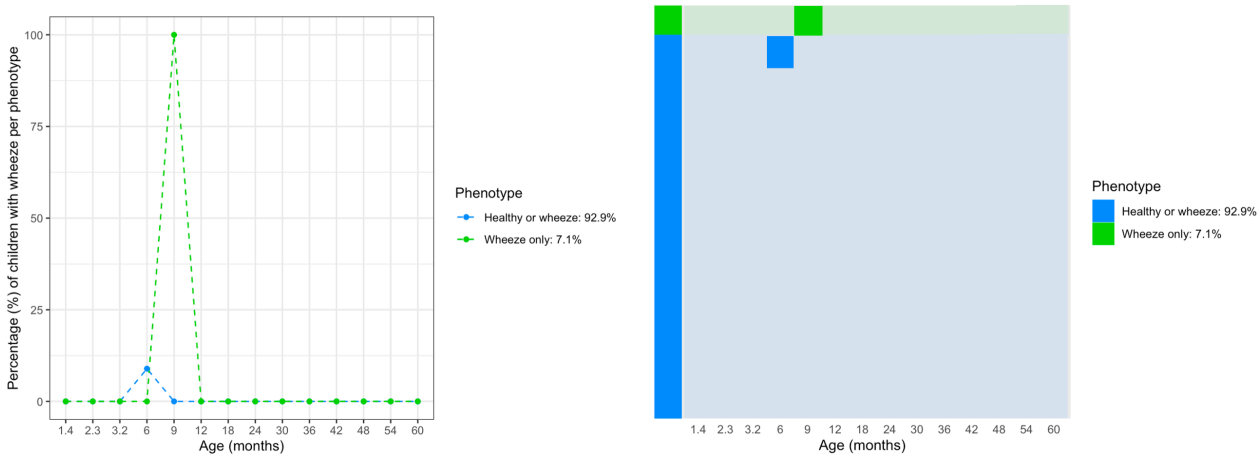


Figure 3.11: A random LCA example with low prevalence: Percentage of children with reported wheeze in the first 5 years of life within each wheeze phenotype (left). Intra-phenotype individual wheeze patterns (right), stratified by phenotype; the darker colours of the bars or lines represent the presence of wheeze and lighter colours represent the absence of wheeze.

In the second scenario, random LCA was applied to a subset of DCHS data that contained never wheezing (00000000000000) and three patterns of wheeze (00010000100000, 00010100000000, 00011010000000), each occurring 4 times. In this case, wheezing occurs at least twice in subjects, and the goal is for a simple 2-class solution to differentiate between healthy and wheezing subjects. It was observed that in this scenario, random LCA perfectly separates healthy and wheezing subjects, Figure 3.12.

Essentially, the limitations of using the LCA approach within the DCHS cohort are thought to stem from the relative probabilities of wheeze prevalence. To summarise further, LCA struggles to differentiate between healthy and wheezing children due to its probabilistic nature, which tends to create classes with varying prevalence rather than distinct groups. The complexity and infrequency of wheezing patterns in the DCHS data, recorded at 14 different time points, result in a large number of possible response patterns, many of which are rare. This data sparsity, combined with the low prevalence of wheezing events, makes it difficult for LCA to form clear distinctions. The observed limitations of LCA in the DCHS context are not a critique of the LCA method itself, but rather a reflection of the specific characteristics of the dataset.

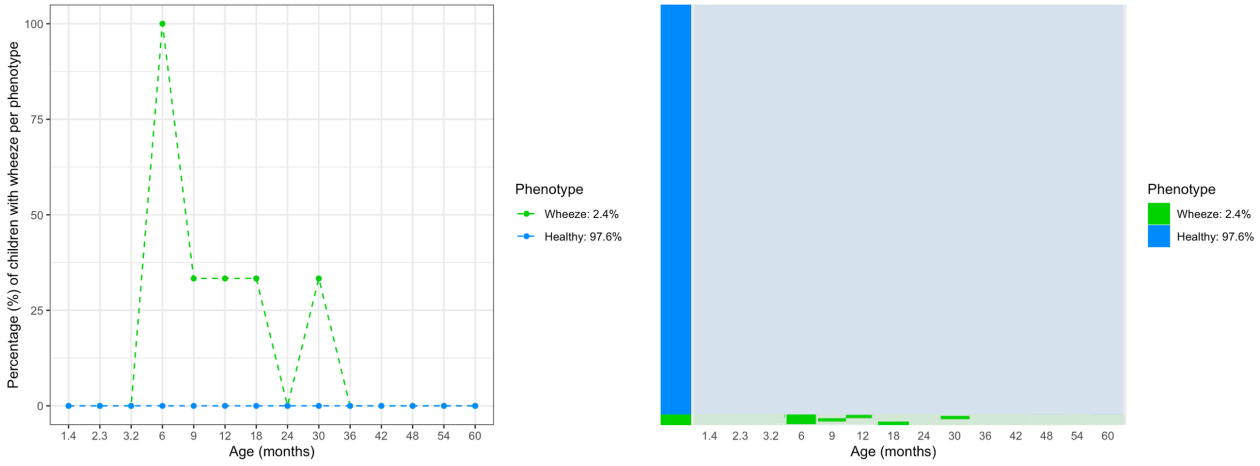


Figure 3.12: A random LCA example with higher prevalence: Percentage of children with reported wheeze in the first 5 years of life within each wheeze phenotype (left). Intra-phenotype individual wheeze patterns (right), stratified by phenotype; the darker colours of the bars or lines represent the presence of wheeze and lighter colours represent the absence of wheeze.

3.8 Phenotype stability

Evaluating the quality of a clustering solution obtained using model-based and traditional distance-based algorithms remains a challenging problem. Unlike supervised learning, there is no absolute truth against which to test the findings formally. However, several methods can be used to quantify and assess the stability of clusters. A broad definition of cluster stability is given by von Luxburg (2010), who defined a solution as stable if different datasets, sampled from the same underlying joint distribution, produce the same clusters, [127]. In this section, a repeated sampling approach is employed to assess the stability of the optimal number of clusters, as outlined in Algorithm 2.

To determine the stability of wheezing phenotypes to changes in the data, multiple iterations of the PAM and LCA models were run while sampling random subsets of children of varying sample size, decreasing in increments of 10% from the full set of children until only half of the children were included. This enables the investigation of both the impact of samples of different sizes and the order of the input data on the clustering results. This process is described through Algorithm 2 using PAM and applied similarly using LCA where the silhouette width is replaced with the BIC.

Algorithm 2 Class stability in the clustering solutions obtained from the PAM model

procedure PHENOTYPE STABILITY
  $\mathbf{N} \leftarrow$ Full set of subjects included in the analysis
  for `sample.size` *in* $[1, 0.9, 0.8, 0.7, 0.6, 0.5]$ **do**
    data.sample $\leftarrow$ Random sample of size $\mathbf{N} *$ `sample.size`
    for `number.phenotypes` *in* $[2:7]$ **do**
      PAM model with classes equal to `number.phenotypes` and data equal to `data.sample`
      Store the silhouette width from the model for each `number.phenotypes`
    end for
  end for
  return The average silhouette width and standard deviation for each phenotype size within `sample.size` across all iterations.
end procedure

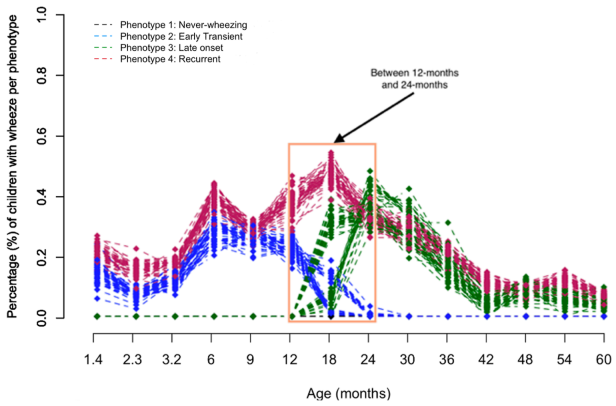
Phenotype Stability - PAM

Analysis of the phenotype stability demonstrated that the clustering solution obtained through the PAM algorithm is stable in that the optimal number of phenotypes does not change with changes in the data, Table 3.9. Through multiple iterations and random subsets of different sizes, it was shown that the same 4 wheezing phenotypes, in terms of characteristics and wheeze prevalence, were consistently identified as the optimal number of phenotypes.

However, repeated data sampling revealed uncertainty about the timing of where the phenotype 2 wheeze profiles end and phenotype 3 wheeze profiles begin within the 12-to-24-month window, Figure

Table 3.9: Wheezing phenotypes from the PAM model: Average silhouette width and standard deviation for subsets across iterations

	Size of subset: silhouette width (std)					
Phenotypes	100%	90%	80%	70%	60%	50%
2	0.63013 (0.000)	0.63152 (0.0052)	0.62988 (0.0062)	0.63143 (0.0060)	0.62995 (0.0096)	0.63227 (0.0157)
3	0.67135 (0.000)	0.67163 (0.0044)	0.67124 (0.0052)	0.67215 (0.0050)	0.67119 (0.0090)	0.67316 (0.0110)
4	0.70955 (0.000)	0.71329 (0.0038)	0.71308 (0.0062)	0.71254 (0.0068)	0.71326 (0.0096)	0.71428 (0.0094)
5	0.67409 (0.000)	0.68706 (0.0251)	0.69265 (0.0289)	0.71263 (0.0328)	0.69964 (0.0321)	0.69867 (0.0271)
6	0.70130 (0.000)	0.70223 (0.0052)	0.70317 (0.0055)	0.70480 (0.0074)	0.70598 (0.0124)	0.70239 (0.0095)
7	0.70233 (0.000)	0.70432 (0.0092)	0.70527 (0.0081)	0.70904 (0.0110)	0.70953 (0.0091)	0.71053 (0.0150)



[1]We ran multiple iterations of the PAM algorithm while sampling random subsets of children of varying sample size with decrements of 10% from the full set of children until only half of the children were included. In each run, indicated by a separate line, 4 phenotypes was the optimal solution.

Figure 3.13: Profiles of wheeze phenotypes over time - Analysis of class stability with respect to changes in different sample sizes[1] of data

3.13. This is illustrated by the change in the shape of the wheezing profiles for phenotypes 2 and 3 across different data subsets. This addresses the question of certainty regarding the boundary between where phenotype 2 stops and where phenotype 3 begins. The answer is that there is understandable uncertainty around where these points lie exactly on the temporal scale.

Multiple implementations of the PAM model show that children assigned to phenotype 2 stopped wheezing between 12 months and 18 months (approximately 70% of the time) or between 18 months and 24 months (approximately 30% of the time). Further results from multiple runs of the model indicated that children assigned to phenotype 3 consistently start wheezing between 12 months and 18 months, but with different levels of wheeze prevalence at 18 months where approximately 30% of the time a low wheeze prevalence was observed at 18 months and approximately 70% of the time a relatively higher wheeze prevalence was observed. Both phenotypes 2 and 3 have similar wheeze characteristics and mainly differ on the temporal scale with an overlap occurring between 12 months and 24 months.

Phenotype Stability - LCA

The results in Table 3.10 show that implementing LCA using random subsets of children of varying sample size led to consistently selecting the two-class model as the optimal model based on the BIC values, indicating high stability and consistency in selecting this model. For the 3-class model, BIC values were slightly higher and showed small variations, with the BIC increasing marginally as the subset size decreased. The standard deviations were minimal, suggesting some consistency but less stability compared to the 2-class model. For the 4 class, 5 class, 6 class, and 7 class models, BIC values were progressively higher and displayed increasing standard deviations as the number of classes increased, indicating less stability and more variability in these models across different subset sizes.

Table 3.10: Wheezing phenotypes from the LCA model: Average BIC and standard deviation for subsets across iterations

	Size of subset: BIC (std)					
Phenotypes	100%	90%	80%	70%	60%	50%
2	5954.243 (0.001)	5954.243 (0.001)	5954.243 (0.001)	5954.243 (0.001)	5954.243 (0.001)	5954.243 (0.001)
3	5998.755 (0.001)	5998.760 (0.014)	5998.802 (0.147)	6001.025 (7.163)	5998.802 (0.147)	5998.764 (0.019)
4	6068.089 (0.001)	6068.562 (1.496)	6068.123 (0.106)	6068.089 (0.001)	6068.122 (0.105)	6068.130 (0.130)
5	6142.467 (1.219)	6141.865 (1.284)	6142.609 (1.908)	6142.463 (0.956)	6141.961 (1.578)	6142.561 (1.769)
6	6215.297 (2.287)	6216.802 (2.850)	6216.274 (2.673)	6215.587 (3.195)	6216.755 (2.914)	6216.806 (4.321)
7	6294.135 (3.242)	6292.974 (3.749)	6292.891 (4.521)	6291.815 (2.251)	6293.641 (4.951)	6293.919 (5.032)

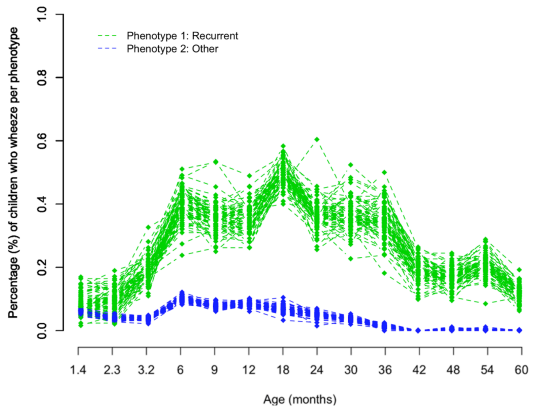


Figure 3.14: Profiles of wheeze phenotypes over time - Analysis of class stability with respect to changes in different sample sizes.

3.9 Comparison of input variables for PAM and LCA

This section examines and discusses the use of PAM and LCA with different types of input data. LCA typically clusters binary indicator data, while PAM usually handles continuous or categorical features. A feature-based PAM approach using mixed variable data was employed in earlier sections of this chapter. Since PAM can also cluster binary data using a Jaccard distance measure, comparing LCA on binary indicators with PAM on features may not provide a fully equivalent analysis. To improve comparability, both methods could be applied to the same type of data, either by adapting PAM for binary data or using LCA on binary indicators.

Table A.2 in the appendix and Figure 3.15 present the clustering results and ASW for PAM applied to binary wheeze data in the DCHS cohort. The focus is on the 4-class clustering solution between both approaches. The 4-class PAM solution in Figure 3.15 is compared to the 4-class LCA solution from Figure 3.9.

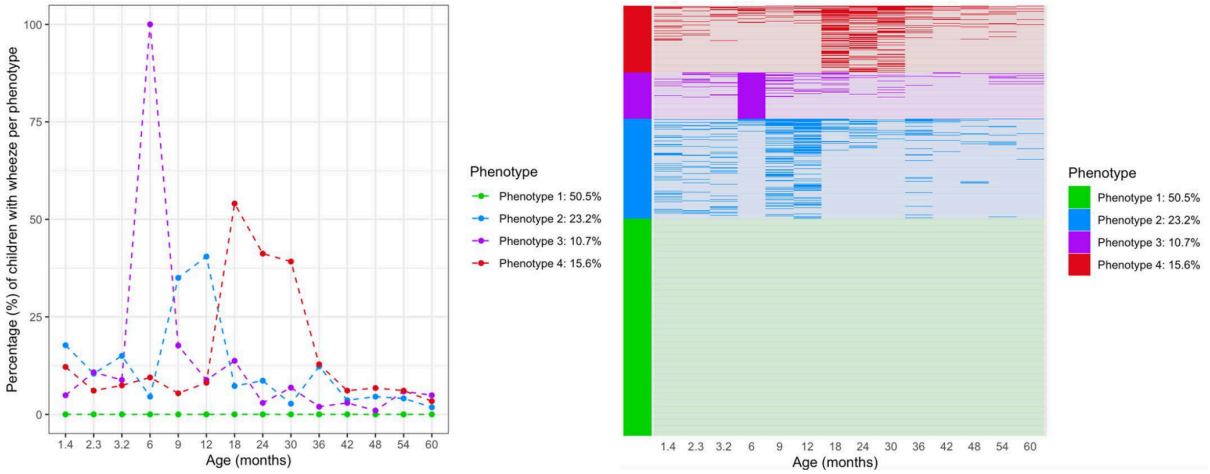


Figure 3.15: Percentage of children with reported wheeze in the first 5 years of life within each wheeze phenotype (left). Intra-phenotype individual wheeze patterns (right). Application of PAM to binary wheezing outcomes.

Phenotype distribution and timing were observed as a difference between the approaches. PAM using binary data appears to identify classes based on higher wheezing prevalence during specific time periods. For example, phenotype 2 in Figure 3.15 represents children who experienced wheezing at 6 months, while phenotype 2 represents children who experienced wheezing between 9 and 12 months, and phenotype 4 represents those who experienced wheezing between 18 and 30 months. While LCA, Figure 3.9, allocates children to phenotype classes based on collective prevalence over the entire observational period. As with prior analysis, PAM applied to binary wheeze data isolated a phenotype class of healthy, non-wheezing children, a result not achieved by the LCA approach.

PAM, when applied to binary data, may lose some interpretability in the class structure due to the limitations of distance metrics on binary features. The Jaccard distance focuses only on the presence or absence of shared features, which can be informative, but it ignores the magnitude or frequency of these features' co-occurrence. Additionally, binary data may not reflect the temporal or sequential relationships that could exist in a longitudinal dataset. This inability to explicitly account for temporal relationships may lead to a less precise identification of patterns.

3.10 Risk factors for wheezing phenotypes identified using the PAM approach

3.10.1 Introduction

The analysis presented in this section focuses on the identification of risk factors for each wheezing phenotype which was identified and described by the 4-class solution using the PAM approach tailored for longitudinal data. As outlined earlier in this chapter, the 4-class solution categorises individuals into the following phenotypes: 1) Never wheeze (480/950, 50.5%), 2) Early transient wheeze (215/950, 22.6%), 3) Late onset wheeze (104/950, 11.0%), and 4) Recurrent wheeze (151/950, 15.9%).

This section of the chapter provides a more detailed discussion of the risk factor analysis that was completed and reported by McCready et al. (2023), [25]. The objective is to describe the characteristics of children in each phenotype. The risk factors chosen for analysis were selected based on evidence from studies conducted on cohorts in HICs as well as associations already published in the DCHS, [24].

The remaining portion of this chapter is structured as follows. (1) The analysis begins by identifying potential risk factors using a Directed Acyclic Graph (DAG) to illustrate their complex interrelationships. (2) Each risk factor is described along with any required encoding for inclusion in the modelling process. (3) The methods underlying the statistical analysis are introduced which investigated associations between wheezing phenotypes and early-life risk factors, followed by the detailed modelling setup implemented. (4) Clinical results are presented in three sections: descriptive tables, model results focusing on child, maternal, and SES characteristics, and model results focusing on lung function. (5) The linearity assumption for quantitative predictors is tested, and comprehensive model diagnostics are performed to validate the approach. (6) Finally, a cross-validation framework is used to confirm the stability and reliability of the selected variables and their corresponding estimates. This structured approach provides a thorough understanding of the risk factors influencing wheezing phenotypes, ensuring the robustness and validity of the findings.

3.10.2 Identifying potential risk factors using a DAG

The relationship between risk factors and the wheezing phenotype outcome, along with potential confounding, is described using a DAG. The DAG in Figure 3.16 displays all the potential risk factors of interest. These factors include exclusive breastfeeding at 6 weeks, preterm birth, the season of birth, sex, antibiotic exposure, LRTI and associated viruses (RSV, RV, AV, Influenza A/B/C, and Parainfluenza 1/2/3), individual SES components (maternal education, income, asset ownership, and household size), maternal depression, maternal allergy, maternal smoking, psychological distress, and IPV.

The DAG serves to illustrate the complex interrelationships between these possible risk factors and

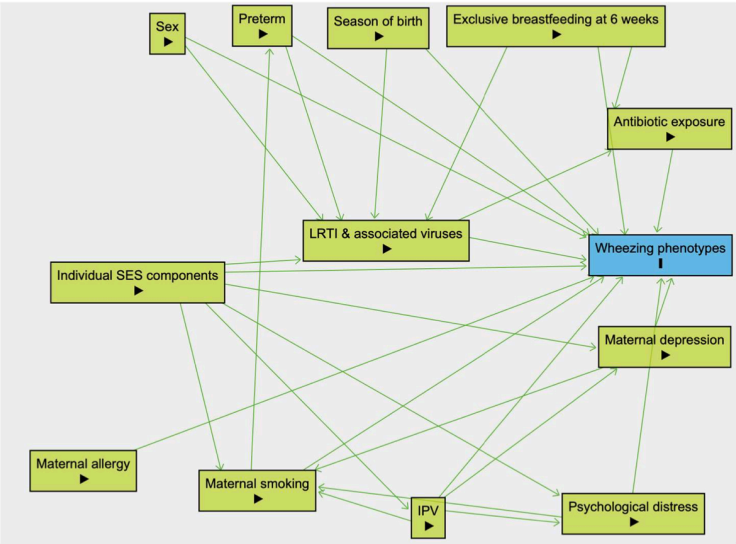


Figure 3.16: Directed acyclic graph (DAG) for the evaluation of risk factors in the multinomial logistic regression model

their potential impacts on the wheezing phenotype outcomes, highlighting how each factor may influence or interact with others. This visualisation aids in understanding the multifaceted nature of the risk factors and helps in identifying which variables may act as confounders, thereby providing a clearer picture of their direct and indirect effects on the wheezing phenotype.

The DAG in Figure 3.16 does not indicate a main exposure and confounders for the association between this exposure and the phenotypes. Some risk factors, such as maternal smoking, act directly upon the wheezing phenotypes, while others, like maternal depression and individual SES components, influence both the phenotypes and other risk factors, thereby becoming potential confounders for each other's association with phenotype.

3.10.3 Descriptions of risk factors and their encoding

Descriptions of risk factors

Each risk factor was measured using specific criteria and methodologies, as defined in Table 1.2. Birth weight z-scores were calculated using the Fenton newborn growth charts for preterm infants and WHO growth charts for term-born infants, with trained clinical staff recording newborn birth weights at delivery using a digital scale. Preterm birth was defined as gestational age at birth less than 37 weeks, estimated using antenatal ultrasound measurements or other methods if unavailable. HIV exposure was determined by testing children born to HIV-positive mothers at multiple intervals using PCR, ELISA, or rapid antibody testing. Exclusive breastfeeding at 6 weeks was reported by

mothers.

Maternal characteristics were assessed through self-reported questionnaires and interviews. Smoking status was evaluated using the ASSIST tool antenatally and postnatally, while maternal asthma or allergy was self-reported antenatally. IPV was assessed using a WHO-adapted questionnaire, and maternal depression was measured using the Edinburgh Postnatal Depression Scale (EPDS) with a cut-off value of 13. Psychological distress was measured using the Self-Reported Questionnaire 20-item (SRQ20) with a cut-off value of 8.

SES components included household income, asset ownership, household size, and maternal education. Income was categorised into three levels based on average monthly household income at maternal enrolment. Asset ownership was scored based on 13 questions about household amenities and services, and household size was categorised into quartiles. Maternal education levels were recorded as primary, some secondary, completed secondary, or any tertiary education.

LRTI was diagnosed based on WHO criteria, and specific viral LRTIs (RSV, RV, AV, Influenza, Parainfluenza) were confirmed via PCR results from nasopharyngeal swabs. An extensive description of the LRTI collection is defined in Table 1.2.

To collect lung function data, airway oscillometry was performed at 6 weeks in unsedated infants during quiet sleep and at 5 years in children sitting comfortably with a nose clip in place, as previously described, [24, 11]. The intra-breath measurements included in the analysis were respiratory resistance at the end of expiration (Ree) and respiratory reactance at the end of expiration (Xee), points of zero flow. Measurements were collected in healthy children at least 4 weeks after a LRTI or wheezing episode.

Encoding of risk factors

The risk factors were recoded to be included in the modelling process. Where repeated LRTI episodes occurred, the number of LRTI episodes per child was treated as a count variable, whereas binary indicators were used for the presence or absence of RSV, RV, adenovirus, influenza, parainfluenza virus, or hospitalised admission at least once during follow-up.

Maternal and child characteristics, except for the season of birth, were encoded as binary variables for modelling. Maternal variables included smoking status, asthma or allergy, IPV, depression, psychological distress, and caesarean delivery, all coded as 1 for presence and 0 for absence. Child variables included sex (1 for male, 0 for female), preterm birth (1 for preterm, 0 for full-term), HIV exposure (1 for exposed uninfected, 0 for unexposed), and exclusive breastfeeding at 6 weeks (1 for exclusively breastfed, 0 for not).

SES variables were encoded as categorical variables for the model. Maternal education was compared against primary education, income against the lowest bracket (<R1 000), asset ownership against the lowest level, and household size against the smallest household (1-4 members). Each SES

variable was encoded to compare different levels against a reference category, enabling the analysis of their effects on the outcome variables.

3.10.4 Statistical analysis methods

Associations between wheezing phenotypes and early-life risk factors

Multinomial logistic regression was used to investigate associations of each phenotype with early-life or environmental risk factors, including viral LRTIs. The risk factors of interest can be broadly grouped under LRTI, maternal characteristics, infant characteristics, and SES. The results are reported as odds ratios (ORs) with 95% CIs.

In the multinomial logistic regression model, the response variable was the wheezing phenotype, with the never wheezing phenotype set as the reference category. For a phenotype Y with K categories, the probability of each category k (for $k = 1, \ldots, K$) relative to the reference category K is given by:

$$\log \frac{P(Y = k)}{P(Y = K)} = \beta_{0k} + \sum_{j=1}^{p} \beta_{jk} X_j, \tag{3.21}$$

where β_{0k} is the intercept and β_{jk} represents the regression coefficients for each predictor X_j in category k. By exponentiating the coefficients β_{jk}, we obtain the ORs for each predictor, representing the relative odds of phenotype k compared to the reference phenotype:

$$\mathrm{OR}_{jk} = \exp\left(\beta_{jk}\right) \tag{3.22}$$

This formulation allows for the interpretation of ORs as the change in odds of belonging to phenotype k (vs. the never wheezing phenotype) for a unit increase in X_j.

The wheezing phenotypes, derived from longitudinal clustering, were treated as a single categorical assignment per child. While the clustering incorporated longitudinal lung function data, the final classification was static, with each child assigned to one phenotype. Therefore, multinomial logistic regression models excluded random effects, as the analysis compared phenotype membership odds without a multilevel structure.

Unadjusted models, containing one risk factor at a time, were initially fitted, followed by the fitting of adjusted models. Unadjusted models were simpler and proved useful for the initial exploration of the associations, while adjusted models allowed to control for potential confounding variables, which could influence the relationship between the variables of interest, and assess the robustness of the associations.

The association between lung function and wheezing phenotypes was investigated by assessing the predictive power of 6-week lung function values for phenotype allocation through a multinomial

logistic model and comparing lung function at 5 years in the wheezing phenotype using linear regression models adjusted for sex and height. The latter could be expressed as follows:

$$Y = \alpha + \sum_{j=1}^{k-1} b_j \cdot \text{Phenotype}_j + b_{\text{Sex}} \cdot \text{Sex} + b_{\text{Height}} \cdot \text{Height} + \epsilon, \tag{3.23}$$

where Y (e.g., Ree and Xee) represents the lung function outcome at 5 years, predicted by wheezing phenotypes, with coefficients b_j for each phenotype relative to the reference, and further adjusted for sex (b_{Sex}) and height (b_{Height}), where α is the intercept and ϵ the error term. Comparisons were then made between 6-week and 5-year lung function by phenotype using cohort-derived z-scores.

Model diagnostics were conducted to assess key assumptions underlying the linear regression models, including linearity of the data, homoscedasticity (constant variance of residuals), and normality of residuals. The linearity assumption underlying multinomial logistic regression analysis for quantitative predictors was confirmed by comparing models, using a likelihood ratios (LR) test.

To compare the included and excluded children in the DCHS, proportions were analysed using the χ^2 test or Fisher's exact test, as appropriate. Mood's Median Test was employed to compare the medians, where appropriate, between the children included in the study and those excluded.

Detailed modelling setup

The modelling process began with a model containing all possible predictors for the wheezing outcome as indicated in the DAG, Figure 3.16, and included 1) all-cause LRTI and 2) viral-specific LRTI (RSV, RV, AV, Influenza, and Parainfluenza) in the model to investigate the associations between LRTI and wheezing phenotypes. The full set of exposures was included as the minimum sufficient set of confounders. There is not one exposure and confounders, but multiple exposures/predictors that can act as confounders for each other.

From a model that included all variables in the DAG, weaker associations were deleted one variable at a time, and the impact of the deletion on the coefficients of other variables was assessed. Weaker associations were those with larger p-values. Successive models were compared after the removal of a weaker predictor using BIC and AIC values. The model-building process continued until no further improvement in model fit was observed while still retaining strong associations. Deletions that would have resulted in a change in regression coefficients for the remaining variables were retained in the model.

3.10.5 Clinical results

Descriptive results and tables

Among the analysed cohort, various maternal and child health factors were observed. A third of mothers smoked antenatally (286 [30%] of 950 mothers), and a similar proportion postnatally (306 [32%] of mothers). Among the mothers, 206 (22%) were HIV-positive; all were receiving antiretroviral therapy. Only two children (<1%) were HIV-positive. Mental health issues were common, with more than a fifth of mothers reporting depression (203 [24%] of 845 mothers), psychological distress (174 [21%] of 846 mothers), or intimate partner violence (292 [34%] of 849 mothers) antenatally. Births were evenly distributed across the seasons, with 150 (16%) of 950 babies born preterm, of whom 101 (67%) were classified as late preterm. Detailed results for these comparisons are shown in Appendix Table A.1.

A total of 980 LRTI episodes were recorded in 459 children (incidence 0.206 episodes per child-year [95% CI 0.194-0.219]), of which, hospital admission was required for 209 LTRIs (21%), Table 3.11. 170 RSV episodes occurred in 156 children (0.036 episodes per child-year [0.031-0.042]), 298 RV episodes occurred in 195 children (0.064 episodes per child-year [0.056-0.071]), 136 AV episodes occurred in 116 children (0.028 episodes per child-year [0.024-0.034]), whereas influenza (0.014 episodes per child-year [0.011-0.017]) and parainfluenza (0.016 episodes per child-year [0.012-0.018]) were less common, Table 3.11. The median age of the first RSV-LRTI was 8.22 months (IQR 3.49-13.79), which was younger than for RV-LRTI (12.26 months [IQR 5.19-23.13]; p<0.0001), AV-LRTI (12.27 months [8.36-16.73]; p<0.0001), influenza (13.41 months [8.68-18.59]; p<0.0001), and parainfluenza (10.31 months [5.51-19.18]; p=0.72). Among children who were admitted to the hospital for treatment, RSV (0.011 episodes per child-year [95% CI 0.008-0.014]) was more frequent than RV (0.008 episodes per child-year [0.006-0.011]) or AV (0.005 episodes per child-year [0.003-0.007]), influenza (0.001 episodes per child-year [0.000-0.003]), and parainfluenza (0.002 episodes per child-year [0.001-0.004]). Repeated episodes of LRTI occurred in 236 (25%) of 950 children; 12 (1%) had repeated infection with RSV, 56 (6%) had repeated infection with RV, and 15 (2%) had repeated infection with AV. Wheezing at the time of LRTI occurred in 339 (35%) of 980 LRTI episodes and was more common among children with RSV-LRTI (93 [55%] of 170 children; p<0.0001) and RV-LRTI (167 [56%] of 298 children; p<0.0001) than AV-LRTI (51 [38%] of 136 children), influenza-LRTI (one [1%] of 67 children), or parainfluenza-LRTI (none).

The distribution of early-life RSV-LRTI, RV-LRTI, AV-LRTI, influenza-LRTI, and parainfluenza-LRTI stratified by wheeze phenotype is shown in Table 3.12. The prevalence of viral LRTIs showed notable variation across the different wheeze phenotypes; LRTI was most commonly associated with recurrent wheezing (125 [83%] of 151 children) compared with other phenotypes (124 [26%] of 480 children with never wheeze; 143 [67%] of 215 children with early transient wheeze; 67 [64%] of 104 children with late-onset wheeze). LRTIs requiring hospital admission were also highest among children with recurrent wheezing (60 [40%] of 151 children) compared with other phenotypes (33 [7%] of 480 children with never wheeze; 44 [20%] of 215 children with early transient wheeze; 19

	Number of children (n)	Number of episodes (n)	Children with repeated episodes (n)	Episodes per child-year (95% CI)	Median age at first episode, months (IQR)
LRTI					
All-cause LRTI	459	980	236	0.206 (0.194−0.219)	9.24 (3.95−18.57)
RSV	156	170	12	0.036 (0.031−0.042)	8.22 (3.49−13.79)
RV	195	298	56	0.064 (0.056−0.071)	12.26 (5.19−23.13)
AV	116	136	15	0.028 (0.024−0.034)	12.27 (8.36−16.73)
Influenza (types A, B, and C)	61	67	6	0.014 (0.011−0.017)	13.41 (8.68−18.59)
Parainfluenza (types 1, 2, 3, and 4)	70	74	4	0.016 (0.012−0.018)	10.31 (5.51−19.18)
LTRI requiring hospital admission					
Any	156	209	33	0.044 (0.038−0.051)	5.16 (1.67−12.85)
RSV	50	51	1	0.011 (0.008−0.014)	3.02 (1.66−10.55)
RV	31	37	5	0.008 (0.006−0.011)	13.25 (5.19−24.66)
AV	22	23	1	0.005 (0.003−0.007)	10.68 (8.88−16.29)
Influenza (types A, B, and C)	7	7	0	0.001 (0.000−0.003)	9.39 (5.51−13.39)
Parainfluenza (types 1, 2, 3, and 4)	10	10	0	0.002 (0.001−0.004)	7.98 (3.61−10.43)
Wheezing at the time of LRTI					
Any	226	339	69	0.071 (0.064−0.079)	8.18 (4.31−14.19)
RV	89	93	4	0.019 (0.016−0.024)	7.36 (3.02−13.22)
RV	127	167	24	0.035 (0.031−0.041)	8.02 (4.75−14.15)
AV	48	51	1	0.011 (0.008−0.014)	12.00 (8.69−16.70)
Influenza (types A, B, and C)	1	1	0	/	/
Parainfluenza (types 1, 2, 3, and 4)	0	0	0	/	/

LRTI = Lower Respiratory Tract Infection; RSV = respiratory syncytial virus; RV = rhinovirus; AV = adenovirus

Table 3.11: Distribution of lower respiratory illness in DCHS children from birth to age 5 years.

	Never wheeze (n=480)	Early transient (n=215)	Late onset (n=104)	Recurrent (n=151)	Total (n=950)
All-cause LRTI	124 (25.8%)	143 (66.5%)	67 (64.4%)	125 (82.8%)	459 (48.3%)
All-cause LRTI requiring hospital admission	33 (6.9%)	44 (20.4%)	19 (18.2%)	60 (39.7%)	156 (16.4%)
RSV-LRTI	20 (4.2%)	57 (25.6%)	22 (15.4%)	57 (35.1%)	156 (16.4%)
RSV-LRTI requiring hospital admission	7 (1.5%)	16 (7.4%)	3 (2.9%)	24 (15.9%)	50 (5.3%)
RV-LRTI	35 (7.3%)	49 (22.8%)	32 (30.7%)	79 (52.3%)	195 (20.5%)
RV-LRTI requiring hospital admission	1 (0.2%)	4 (1.9%)	9 (8.6%)	17 (11.2%)	31 (3.2%)
AV-LRTI	19 (3.9%)	31 (14.4%)	21 (20.2%)	45 (29.8%)	116 (12.2%)
AV-LRTI requiring hospital admission	2 (0.4%)	6 (2.8%)	3 (2.9%)	11 (7.3%)	22 (2.3%)
Influenza (A or B or C)-LRTI	11 (2.3%)	20 (9.3%)	12 (11.5%)	18 (11.9%)	61 (6.4%)
Influenza (A or B or C)-LRTI requiring hospital admission	1 (0.2%)	2 (0.9%)	1 (0.9%)	3 (2.0%)	7 (0.7%)
Parainfluenza (1 or 2 or 3)-LRTI	12 (2.5%)	17 (7.9%)	10 (9.6%)	31 (20.5%)	70 (7.4%)
Parainfluenza (1 or 2 or 3)-LRTI requiring hospital admission	0 (0.0%)	1 (0.4%)	2 (1.9%)	7 (4.6%)	10 (1.1%)

Children could have more than one type of LRTI; LTRI categories were not mutually exclusive.
LRTI = Lower Respiratory Tract Infection; RSV = respiratory syncytial virus; RV = rhinovirus; AV = adenovirus

Table 3.12: Prevalence of RSV, RV, AV, influenza, or parainfluenza-LRTI, by wheezing phenotype.

[18%] of 104 with late-onset wheeze). Among hospitalised cases, RSV was the most common viral LRTI among children with recurrent wheezing.

Model results focusing on child, maternal, and SES characteristics

In this section, the unadjusted and adjusted associations between early-life factors and wheezing phenotypes are examined. Table 3.13 displays the results of the unadjusted multinomial logistic regression analysis and Table 3.14 displays the results of the adjusted multinomial logistic regression analysis, providing insights into the relationship between various factors and different wheezing phenotypes.

Upon reviewing the unadjusted associations, maternal and infant characteristics, socioeconomic status, all-cause LRTI, and specific viral LRTI were all associated with wheezing phenotypes, Table 3.13. RSV demonstrated strong associations with wheezing phenotypes, highlighting the importance of viral etiology in respiratory outcomes. Not all strong univariate associations entered the adjusted model after the model-building process, Table 3.13.

Postnatal and antenatal maternal characteristics showed a strong correlation, with a stronger association between postnatal exposures and wheeze phenotypes compared to antenatal exposures and wheeze phenotypes. Therefore, postnatal exposures were included in the adjusted model, as displayed in Table 3.14. The associations in the univariate models between LRTI and wheeze phenotypes persisted in the multivariate model but with smaller effect sizes.

LRTI and RSV-LRTI were associated with all three wheezing phenotypes, whereas other respiratory viruses were not strongly associated with any phenotype. Exclusive breastfeeding was protective, especially for early transient wheeze. In addition to all-cause LRTI and respiratory syncytial virus-LRTI, maternal smoking, intimate partner violence, higher socioeconomic status, and male sex were associated with recurrent wheezing, Table 3.14.

The persistence of strong associations for certain factors, such as RSV-LRTI and postnatal smoking, in the adjusted model results, highlights their critical role in the development of wheezing phenotypes.

Table 3.13: Unadjusted multinomial logistic regression of the association of early-life factors with wheezing phenotypes in the DCHS (reference class: Never wheezing)

	Unadjusted Effects					
Phenotype Comparison	Phenotype 2: Early transient		Phenotype 3: Late onset		Phenotype 4: Recurrent	
	OR (95% CI)	p-value	OR (95% CI)	p-value	OR (95% CI)	p-value
LRTI						
Number of LRTI episodes	2.59 (2.13, 3.15)	< 0.0001	2.40 (1.91, 3.02)	< 0.0001	4.06 (3.29, 5.02)	< 0.0001
Hospital admission (vs no hospital admission)	1.15 (0.67, 1.98)	0.61	1.02 (0.50, 2.08)	0.95	2.26 (1.31, 3.88)	0.081
RSV-LRTI (vs RSV negative)	3.26 (1.80, 5.93)	< 0.0001	2.12 (1.35, 4.61)	< 0.029	4.10 (2.23, 7.55)	< 0.0001
RV-LRTI (vs RV negative)	1.22 (0.69, 2.14)	0.48	1.81 (0.93, 3.48)	0.076	1.64 (0.91, 35.01)	0.11
AV-LRTI (vs AV negative)	1.03 (0.54, 1.96)	0.93	1.68 (0.82, 3.42)	0.15	1.12 (0.57, 2.21)	0.72
Influenza-LRTI (A or B or C) vs influenza negative	1.17 (0.52, 2.59)	0.69	1.48 (0.61, 3.59)	0.38	0.58 (0.23, 1.46)	0.25
Parainfluenza-LRTI (1 or 2 or 3) vs parainfluenza negative	1.19 (0.52, 2.68)	0.67	1.18 (0.46, 3.02)	0.73	1.24 (0.54, 2.86)	0.61
Maternal Characteristics						
Antenatal Smoking	1.22 (0.85, 1.74)	0.28	1.53 (0.98, 2.39)	0.059	1.75 (1.19, 2.56)	0.0042
Postnatal Smoking	1.16 (0.82, 1.65)	0.39	1.70 (1.09, 2.65)	0.021	1.97 (1.33, 2.87)	0.0012
Maternal Asthma or Allergy	1.46 (0.73, 2.93)	0.28	1.69 (0.73, 3.92)	0.22	1.57 (0.74, 3.33)	0.23
Antenatal Depression	0.72 (0.47, 1.10)	0.13	0.89 (0.52, 1.41)	0.65	1.18 (0.77, 1.80)	0.46
Postnatal Depression	0.97 (0.66, 1.43)	0.87	0.91 (0.54, 1.52)	0.72	1.38 (0.92, 2.08)	0.12
Antenatal Psychological Distress	1.22 (0.79, 1.87)	0.36	1.03 (0.57, 1.84)	0.91	1.93 (1.25, 2.99)	0.0033
Postnatal Psychological Distress	1.08 (0.66, 1.77)	0.76	1.18 (0.63, 2.22)	0.611	2.01 (1.24, 3.26)	0.0067
Antenatal IPV	0.73 (0.50, 1.06)	0.11	0.87 (0.54, 1.40)	0.56	1.35 (0.92, 1.99)	0.11
Postnatal IPV	1.26 (0.89, 1.77)	0.190	1.68 (1.08, 2.62)	0.019	2.40 (1.64, 3.51)	< 0.0001
Mode of delivery (Caesarean)	1.20 (0.81, 1.80)	0.37	1.25 (0.76, 2.09)	0.37	0.80 (0.49, 1.31)	0.38
Infant Characteristics						
Sex (male vs female)	1.24 (0.90, 1.72)	0.17	1.38 (0.90, 2.17)	0.13	2.64 (1.79, 3.89)	< 0.0001
Pre-term (<37 vs >=37 weeks)	0.90 (0.49, 1.31)	0.65	1.21 (0.68, 2.13)	0.51	1.55 (0.90, 2.35)	0.059
HIV (exposed uninfected vs unexposed)	0.92 (0.63, 1.35)	0.68	1.35 (0.79, 2.32)	0.26	1.45 (0.91, 2.33)	0.12
Exclusive Breast Feeding at 6 weeks	0.72 (0.51, 1.00)	0.051	0.77 (0.50, 1.19)	0.24	0.86 (0.59, 1.25)	0.45
Season of Birth						
Autumn vs Summer	1.70 (1.09, 2.66)	0.023	1.30 (0.72, 2.37)	0.38	1.69 (1.02, 2.81)	0.038
Winter vs Summer	0.84 (0.54, 1.32)	0.46	0.66 (0.35, 1.22)	0.18	0.80 (0.47, 1.34)	0.39
Spring vs Summer	0.82 (0.51, 1.33)	0.43	1.17 (0.65, 2.09)	0.61	0.91 (0.53, 1.56)	0.73
Antibiotic exposure	1.05 (0.61, 1.82)	0.85	1.68 (0.80, 3.53)	0.17	0.72 (0.43, 1.19)	0.21
Socio Economic Status (SES)						
Maternal education						
Tertiary vs primary	1.40 (0.60, 3.25)	0.42	1.76 (0.36, 8.51)	0.48	0.47 (0.15, 1.46)	0.19
Completed secondary vs primary	0.96 (0.50, 1.85)	0.91	3.26 (0.95, 11.14)	0.059	0.85 (0.42, 2.17)	0.66
Secondary vs primary	1.04 (0.55, 1.94)	0.89	2.66 (0.79, 8.93)	0.11	0.83 (0.42, 1.61)	0.58
Income						
R1 000 to R5 000 vs <R1 000	1.17 (0.83, 1.65)	0.36	0.83 (0.53, 1.31)	0.43	0.97 (0.65, 1.45)	0.91
More than R5 000 vs <R1 000	0.93 (0.57, 1.72)	0.98	0.81 (0.39, 1.67)	0.58	1.41 (0.80, 2.47)	0.23
Asset Ownership						
Low-Medium vs Low	1.21 (0.79, 1.85)	0.35	1.90 (0.99, 3.61)	0.051	1.10 (0.68, 1.79)	0.68
Medium-High vs Low	0.78 (0.49, 1.25)	0.31	1.70 (0.87, 3.32)	0.12	0.76 (0.44, 1.30)	0.31
High vs Low	1.17 (0.72, 1.90)	0.51	2.77 (1.42, 5.54)	0.0031	1.30 (0.76, 2.21)	0.33
Household Size						
Small-Medium [4-5) vs Small [1-4)	0.89 (0.56, 1.43)	0.65	1.39 (0.74, 2.58)	0.29	1.12 (0.65, 1.92)	0.67
Medium-Large [5-7) vs Small	0.79 (0.51, 1.20)	0.27	1.06 (0.58, 1.91)	0.84	1.19 (0.74, 1.92)	0.45
Large vs Small [7-18]	1.12 (0.72, 1.75)	0.59	1.79 (1.00, 3.21)	0.051	1.21 (0.72, 2.04)	0.46
Lung function (at 6-weeks)[1]						
Ree (hPa.s.L^{-1})	0.99 (0.97, 1.02)	0.65	1.02 (0.99, 1.04)	0.13	1.02 (0.99, 1.04)	0.081
Xee (hPa.s.L^{-1})	0.99 (0.96, 1.03)	0.76	0.98 (0.94, 1.02)	0.31	0.94 (0.91, 0.97)	0.0021

LRTI = Lower Respiratory Tract Infection; IPV = intimate partner violence, RSV = Respiratory Syncytial Virus; RV = Rhinoviruses; AV = adenovirus; Ree = Respiratory resistance at the end of expiration; Xee = Respiratory reactance at the end of expiration

[1] Lung function at 6-weeks was adjusted for height, sex, and ancestry.

Table 3.14: Adjusted multinomial logistic regression of factors associated with wheezing phenotypes in DCHS (reference class: Never wheezing)

	Adjusted Effects					
Phenotype Comparison	Phenotype 2: Early transient		Phenotype 3: Late onset		Phenotype 4: Recurrent	
	OR (95% CI)	p-value	OR (95% CI)	p-value	OR (95% CI)	p-value
LRTI						
Number of LRTI episodes	1.69 (1.27, 2.30)	0.00053	1.60 (1.13, 2.28)	0.0081	2.79 (2.05, 3.81)	< 0.0001
Hospital admission (vs no hospital admission)	1.40 (0.43, 1.54)	0.52	0.84 (0.38, 1.82)	0.66	1.01 (0.51, 2.02)	0.97
RSV-LRTI (vs RSV negative)	3.34 (1.82, 6.26)	< 0.001	2.82 (1.32, 6.01)	0.0071	2.59 (1.30, 5.15)	0.0067
RV-LRTI (vs RV negative)	1.39 (0.74, 2.59)	0.31	1.79 (0.87, 3.73)	0.12	1.64 (0.83, 3.22)	0.15
AV-LRTI (vs AV negative)	0.87 (0.42, 1.77)	0.71	1.56 (0.73, 3.43)	0.21	1.11 (0.53, 2.30)	0.78
Influenza-LRTI (A or B or C) vs influenza negative	1.36 (0.46, 2.77)	0.78	1.38 (0.51, 3.73)	0.52	0.74 (0.27, 1.95)	0.54
Parainfluenza-LRTI (1 or 2 or 3) vs parainfluenza negative	1.05 (0.41, 2.62)	0.92	1.75 (0.63, 4.81)	0.27	1.15 (0.45, 2.92)	0.76
Maternal Characteristics						
Postnatal Smoking	1.14 (0.75, 1.74)	0.52	1.65 (1.01, 2.73)	0.048	1.88 (1.12, 3.02)	0.015
Maternal Asthma or Allergy	2.13 (1.01, 4.49)	0.047	2.02 (0.81, 4.99)	0.12	2.43 (0.94, 6.26)	0.066
IPV	1.23 (0.83, 1.83)	0.29	1.71 (1.04, 2.80)	0.031	2.01 (1.23, 3.29)	0.0051
Infant Characteristics						
Sex (male vs female)	1.29 (0.88, 1.87)	0.19	1.44 (0.89, 2.33)	0.13	2.47 (1.50, 4.04)	0.00033
Exclusive Breastfeeding at 6 weeks	0.65 (0.45, 0.97)	0.035	0.82 (0.50, 1.32)	0.42	0.72 (0.44, 1.16)	0.18
Socio Economic Status (SES)						
Asset ownership						
Low-medium vs Low	1.31 (0.80, 2.16)	0.28	2.51 (1.19, 5.27)	0.016	1.22 (0.63, 2.31)	0.55
Medium-high vs Low	0.86 (0.50, 1.50)	0.62	2.12 (0.97, 4.63)	0.057	0.89 (0.43, 1.82)	0.75
High vs Low	1.40 (0.79, 2.46)	0.24	4.74 (2.20, 10.21)	< 0.0001	2.46 (1.23, 4.91)	0.017

LRTI = Lower Respiratory Tract Infection; IPV = intimate partner violence; RSV = respiratory syncytial virus;
RV = rhinovirus; AV = adenovirus OR (95% CI) = odds ratio (95% confidence interval)

Model results focusing on lung function

Lung function assessments were highly successful, with 966 out of 981 children (98.5%) having at least one lung function assessment completed by year 5, Figure 1.1. The aim was to minimise the exclusion of children as much as possible. Each measurement contributes to estimating the mean levels at a specific age. Assuming the missing data is random, these single measurements do not negatively impact the estimation of time profiles.

Compared with children who never wheezed, Xee was lower at 6 weeks in children with recurrent wheezing, but not transient or late-onset wheezing, Table 3.13. Ree at 6 weeks was not associated with any phenotype. At 5 years, Ree was higher among children with early transient wheeze (average increase 0.57 $hPasL^{-1}$ [95% CI 0.08-1.06]) or recurrent wheeze (average increase 0.79 $hPasL^{-1}$ [95% CI 0.22-1.36]) than children who never wheezed, Table 3.15. Analysis of residuals indicated no deviations from underlying linear model assumptions, Figure 3.19.

Furthermore, airway resistance exhibited a greater increase between infancy and age 5 years in the recurrent wheeze group compared to other phenotypes. This pattern may potentially predispose these children to impaired respiratory health in adulthood, Figure 3.17. In healthy individuals, z-scores are expected to remain relatively stable over time, as reflected by non-significant changes between time points. However, there was an observed increase in z-score by 5 years in the recurrent wheeze group, indicative of worsening airway narrowing.

Table 3.15: Association of lung function (oscillometry) at 5 years and wheeze phenotypes in DCHS.

	Ree ($hPa.s.L^{-1}$)		Xee ($hPa.s.L^{-1}$)	
	Regression coefficient (95% CI)	p-value	Regression coefficient (95% CI)	p-value
Early transient vs Never-wheeze	0.57 (0.08, 1.06)	0.026	-0.09 (-0.36, 0.18)	0.512
Late onset vs Never-wheeze	0.34 (-0.30, 0.97)	0.291	-0.18 (-0.44, 0.16)	0.298
Recurrent vs Never-wheeze	0.79 (0.22, 1.36)	0.005	-0.23 (-0.55, 0.09)	0.151

Ree = Respiratory resistance at the end of expiration; Xee = Respiratory reactance at the end of expiration
95% CI = 95% confidence interval
Linear regression models adjusted for sex and height.

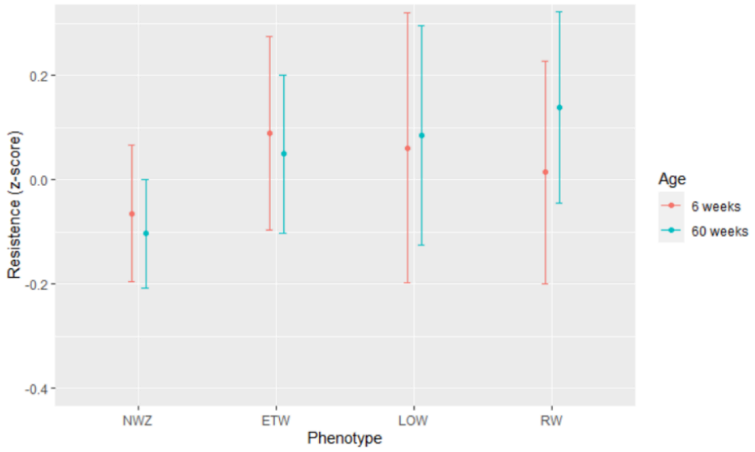


Figure 3.17: Airway resistance at ages 6 weeks and 5 years in four wheeze clusters in DCHS (mean z-scores, 95% CI). NWZ = Never wheeze; ETW = Early transient wheeze; LOW = Late onset wheeze; RW = Recurrent wheeze

Linearity assumption

The linearity assumption underlying multinomial logistic regression analysis for quantitative predictors was confirmed by comparing models, using a likelihood ratios (LR) test, with the continuous lung function predictors to models with a categorical version of the predictor divided into percentiles (LR p-value = 0.51 for Ree; and LR p-value = 0.23) and further compared to splines models for B-spline basis matrix for a polynomial spline with 3 (LR p-value = 0.046 for Ree; and LR p-value = 0.29) and 5 (LR p-value = 0.087 for Ree; and LR p-value = 0.078) degrees of freedom. The results suggest that no deviation from linearity was observed. The p-values from the LR tests indicate that the continuous predictors fit the model well without needing to be transformed into categorical predictors or splines, except for a slight indication of non-linearity for Ree in the 3 degrees of freedom spline model. The linearity assumption underlying multinomial logistic regression analysis for quantitative predictors can also be tested by plotting the probability of a wheezing phenotype versus the continuous lung function outcome (Xee and Ree) for a linear model (blue), cubic spline with 3 df (black), and lung function as a categorical variable divided into percentiles (red), Figure 3.18. The strong overlap between the linear model (blue), cubic spline with 3 df (black), and categorical percentiles (red) indicates no clear deviation from linearity for the lung function predictors in the multinomial logistic regression model.

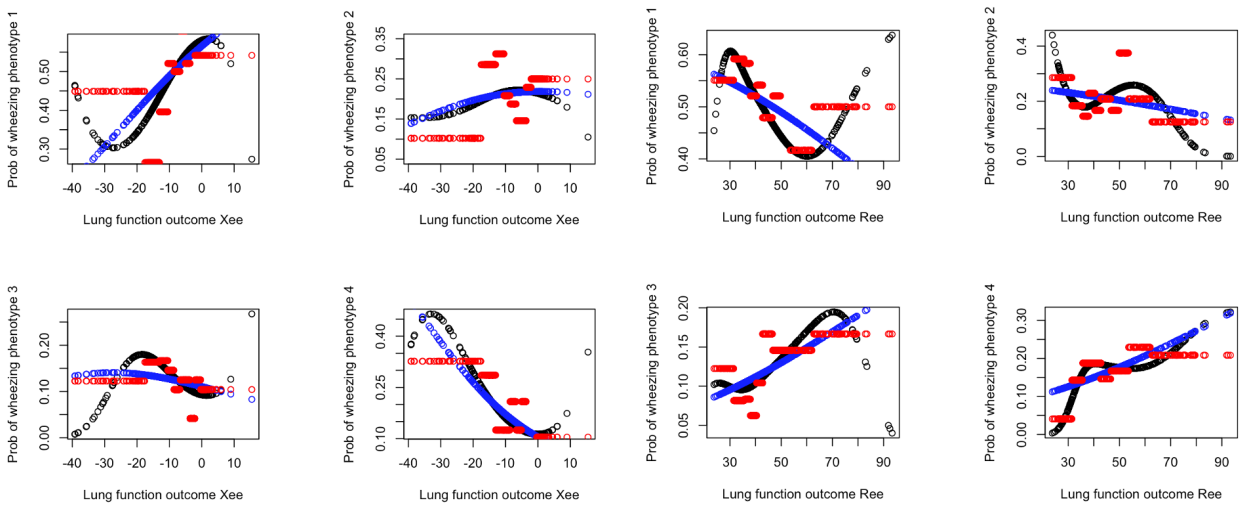


Figure 3.18: Probability of wheezing for a phenotype versus the continuous Xee outcome (left) and Ree (right).

Model diagnostics

Model diagnostics were used to assess the assumptions underlying the linear regression models. The residual (versus fitted) plot showed no fitted pattern. The homogeneity of variance of the residuals

was assessed by plotting the square root of the standardised residuals against the fitted values; a random scatter was observed. QQ plots showed no clear deviations from normality. Analysis of residuals indicated no deviations from underlying linear model assumptions, Figure 3.19.

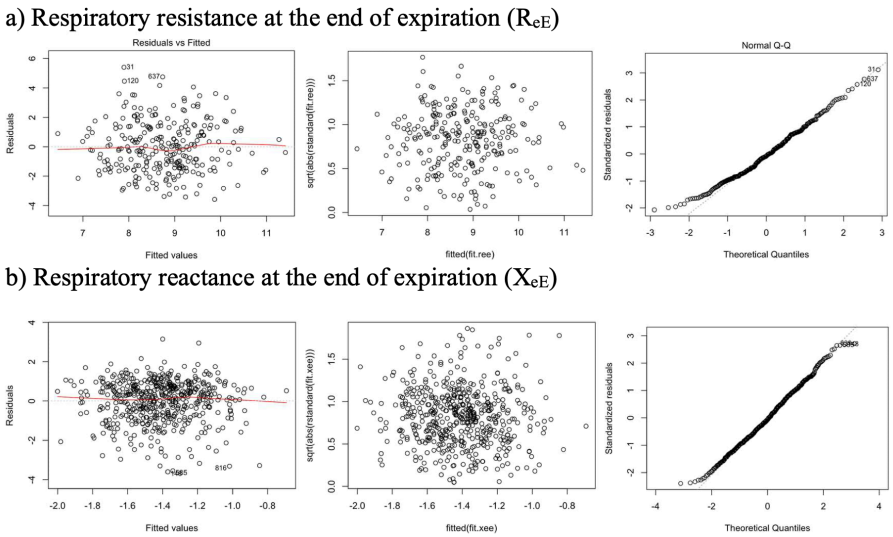


Figure 3.19: Analysis of residuals plots assessing linearity of data (left), homogeneity of residuals variance (centre), and normality of residuals (right)

3.10.6 Cross validation framework for variable selection and stability

Background

Cross-validation (CV) is used in this section as a resampling method as it involves fitting the same statistical method multiple times using different subsets of the data, providing an opportunity to evaluate a model's performance and stability. In this study, CV was used to perform two functions. Firstly, for variable selection to a multivariate model, and secondly to determine the stability of the model estimates. Relying on a single dataset for variable selection can result in instability, where small changes in the data, such as adding or removing a few data points, can lead to different variables being selected, indicating a lack of robustness in the process. Additionally, the model may struggle to generalise to other datasets because it might capture patterns unique to the specific dataset rather than universal relationships, reducing its ability to perform well on unseen data. The CV approach introduced here is not a replacement for the modelling setup introduced earlier in this chapter and employed by McCready et al. (2023), but rather acts as a method to confirm the prior results and their robustness.

Procedure

The framework is structured as a two-step iterative process. Initially, the dataset comprising N observations is divided into k subsets, with each subset varying in size from 100% to 50% of the original dataset. This random sampling is repeated k times, ensuring a diverse range of data scenarios. In the first step of the process, variables are selected using a stepwise model selection approach, specifically employing the stepAIC() function in R, which considers both forward and backward selection methods based on the Akaike Information Criterion (AIC). The selected variables from each subset are then stored in an array. The goal is to identify the most relevant variables based on the proportion of times each variable was selected (using stepwise model selection) in the CV calls, Figure 3.20 (step 1). As a soft indicator, variables were selected if they occurred in approximately 50% of the CV calls. This framework was used to enhance confidence in the main risk factors identified in previous stages of the analysis and to assess the stability of variable selection across different subsamples.

In the second step, the CV framework provides insights into the stability of the model estimates across the varied dataset samples, Figure 3.20 (step 2). By examining the variability in the odds ratios and p-values obtained from the output during the CV procedure, the study demonstrated how changes in the input dataset, such as sample size, can lead to variability in the model output. By doing so, this approach attempts to provide more oversight into the strength and consistency of the associations between the wheezing phenotypes and the predictors thereof. This can be particularly helpful in identifying false or weaker associations which are not regularly observed across identical models using slightly different subsamples of data as input. If an association is consistently observed across various models with different data subsets, it adds more weight to its validity. This rigorous approach can help researchers identify which associations are robust and which might be spurious.

This comprehensive CV framework thus not only helps in identifying the most influential variables but also in verifying the consistency of the model estimates under different data conditions.

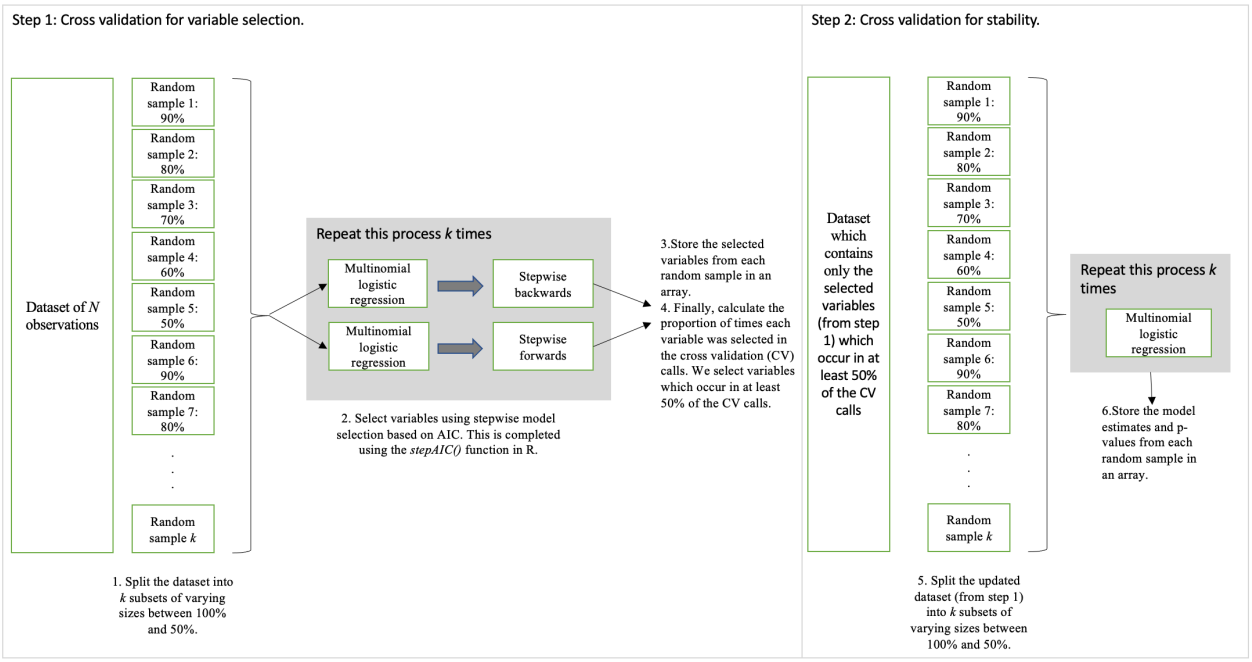


Figure 3.20: A cross-validation (CV) framework for variable selection and stability

Results

Step 1 of the CV approach was applied using the previously identified wheezing phenotypes and their associated risk factors under the PAM approach, Figure 3.21. It was found that LRTI was repeatedly identified in every iteration of the CV framework. RSV-LRTI and sex were both identified in most, but not all, of the iterations. Other frequently selected risk factors included maternal postnatal smoking, IPV, and asset ownership, while SES factors were occasionally selected. The results mostly confirm the final model introduced earlier in Table 3.14, with the exception of exclusive beast feeding at 6 weeks and maternal asthma or allergy. All-cause LRTI and 2) viral-specific LRTI (RSV, RV, AV, Influenza, and Parainfluenza) would always be included in the model due to their high clinical relevance in the DCHS cohort.

Figure 3.22 displays the stability of the odds ratios and p-values obtained from the multinomial model output during the CV procedure under step 2. All data points are displayed on top of the box plots (including outliers). These figures indicate how changes in the input dataset lead to variability in the model output. Greater changes in the input data, indicated by smaller sample sizes which are highlighted by the lighter blue points, lead to increased variability, Figure 3.22. These findings align

with the idea that smaller sample sizes typically result in less precise estimates, thereby affecting the reliability of p-values, [128, 129]. Even within a fixed sample size, there was some variability in the exact estimate of the OR, which is expected due to the nature of resampling. This variability reflects the inherent uncertainty in the estimate, and it is captured by the range of values within a confidence interval. A 95% CI for the OR is designed to encompass this variability. Therefore, the 95% CI's for the OR should cover the variability within fixed sample sizes.

The consistency and robustness of specific associations become evident in Figure 3.22. A good example of this is the association between wheezing phenotypes and sex (male versus female). It was previously reported in Table 3.14 that male sex was associated with recurrent wheezing (2.47 [95% CI 1.504.04], p=0.00033). Figure 3.22 (g)-(h) confirms the strength of this association with recurrent wheeze through odds ratios which are consistently greater than 1 and p-values which are consistently below 0.05 with the exception of a few model estimates where the sample size was 50% of the total sample of 950. Other strong and robust associations with recurrent wheezing include LRTI and RSV-LRTI. Previously, a strong association was found between RSV-LRTI (2.82 [95% CI 1.32- 6.01], p = 0.0071) and late onset wheezing. However, the stability analysis in Figure 3.22 (o)-(p) reveals that this association is not robust, as a strong association is not consistently observed given smaller data samples. This is a result of the lower statistical power associated with smaller sample sizes. To some extent, the analysis highlights the fundamental concept of power.

The 95% CI of the model estimates (ORs) are not directly shown in the accompanying figures within this section for each iteration of the CV procedure. Although, an underlying distribution can be assumed within a sample. The 95% CI interval is therefore the range in which we would expect the true parameter to lie in 95% of repeated samples, given the same underlying data-generating process. The density plots of the OR estimates by sample size illustrate this concept and are shown in Figure 3.23. As the sample size increases (from 50% to 90%), there is a general trend of increasing data density, which indicates that smaller intervals are associated with larger samples. This trend underscores the principle that larger sample sizes tend to produce more precise estimates, as reflected by narrower confidence intervals. This is because the effect of random variability within the sample is reduced.

What these results indicate is that there are two areas of variability. 1) Variability within data samples of the same sample size. 2) Variability between data samples of different sizes (for example, samples of 50% versus 80%).

To summarise, this CV approach ensures that the selected variables and their corresponding estimates are not only strongly associated with the outcome of interest, but also consistently reliable across various subsampling scenarios. The results confirm the robustness of key associations, such as those between wheezing phenotypes and specific risk factors like LRTI and sex, while highlighting the variability introduced by different sample sizes. While there is variability, the general conclusions are validated.

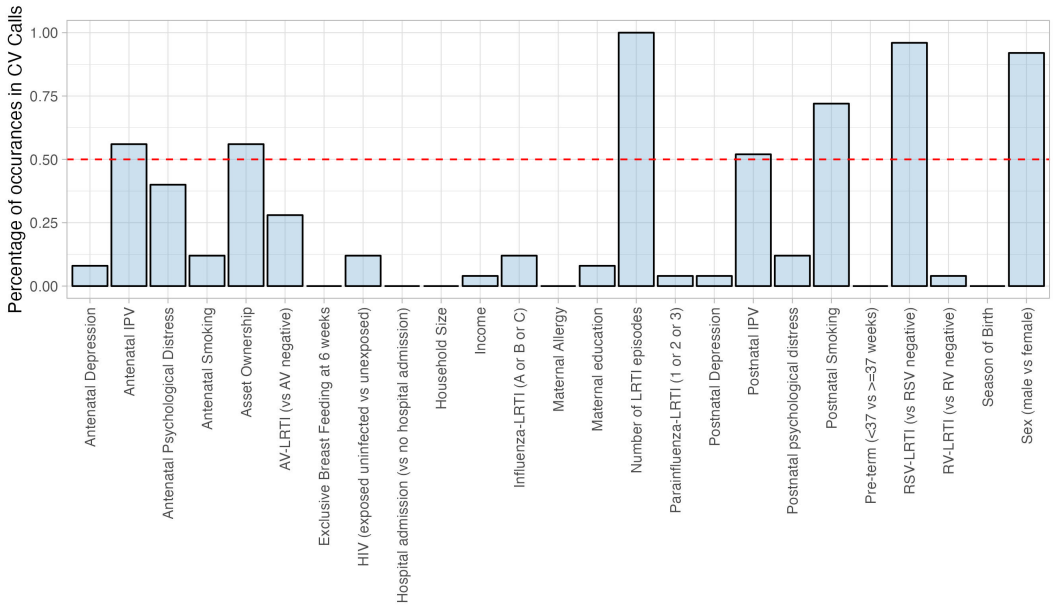


Figure 3.21: Cross-validation for variable selection

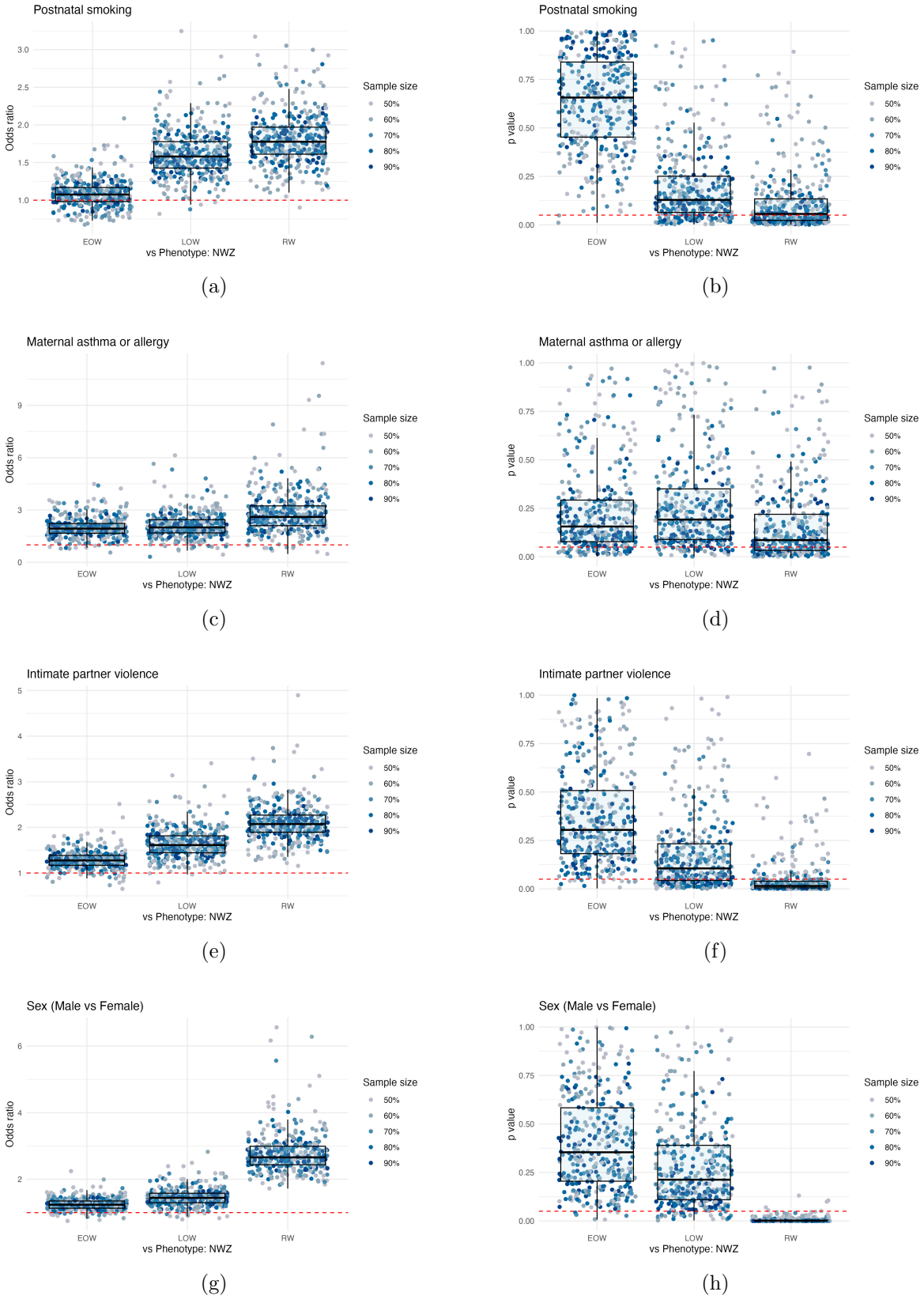


Figure 3.22: Stability analysis sex using cross-validation. Odds ratios (left) and p-values (right).

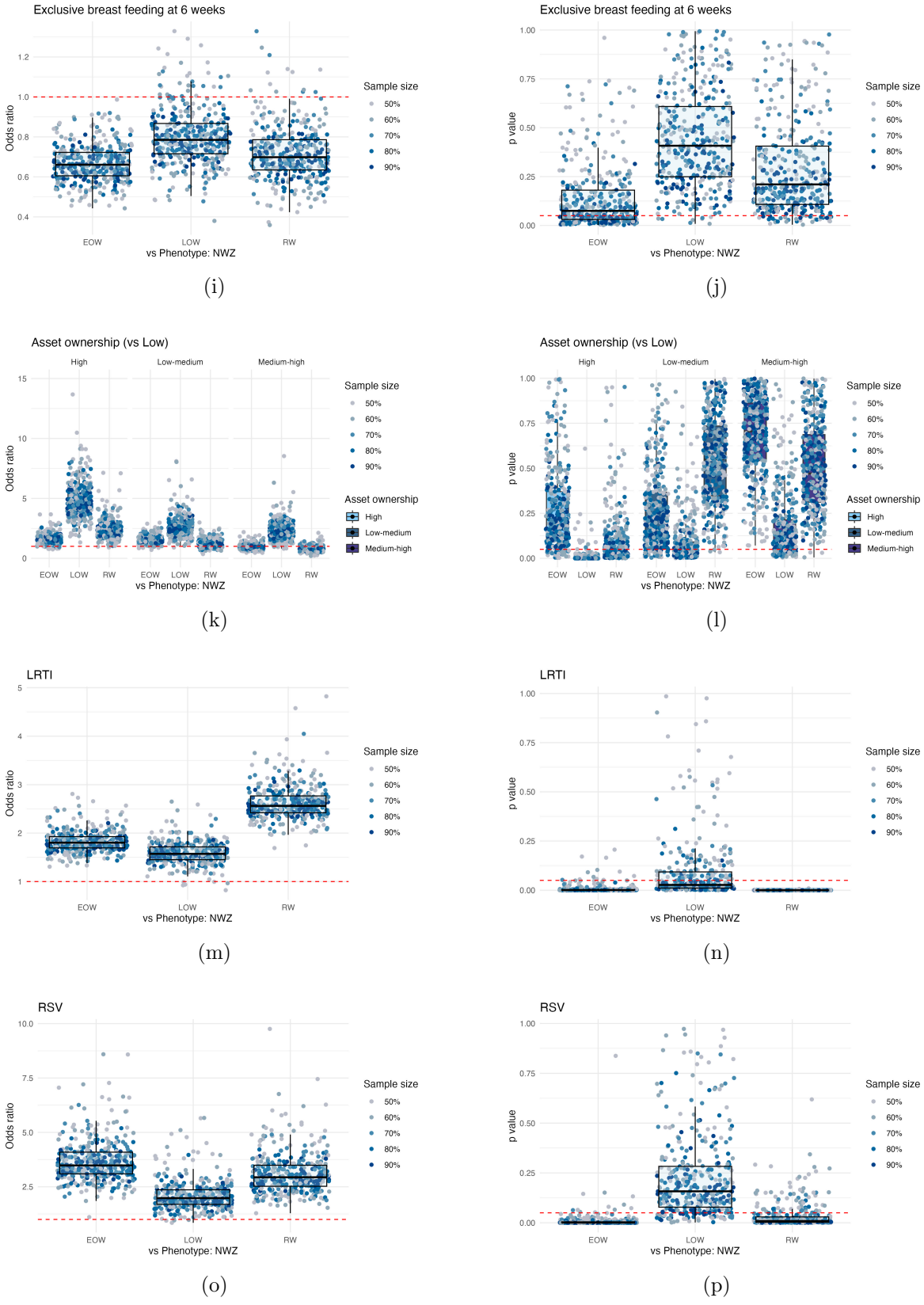


Figure 3.22: Stability analysis sex using cross-validation. Odds ratios (left) and p-values (right).

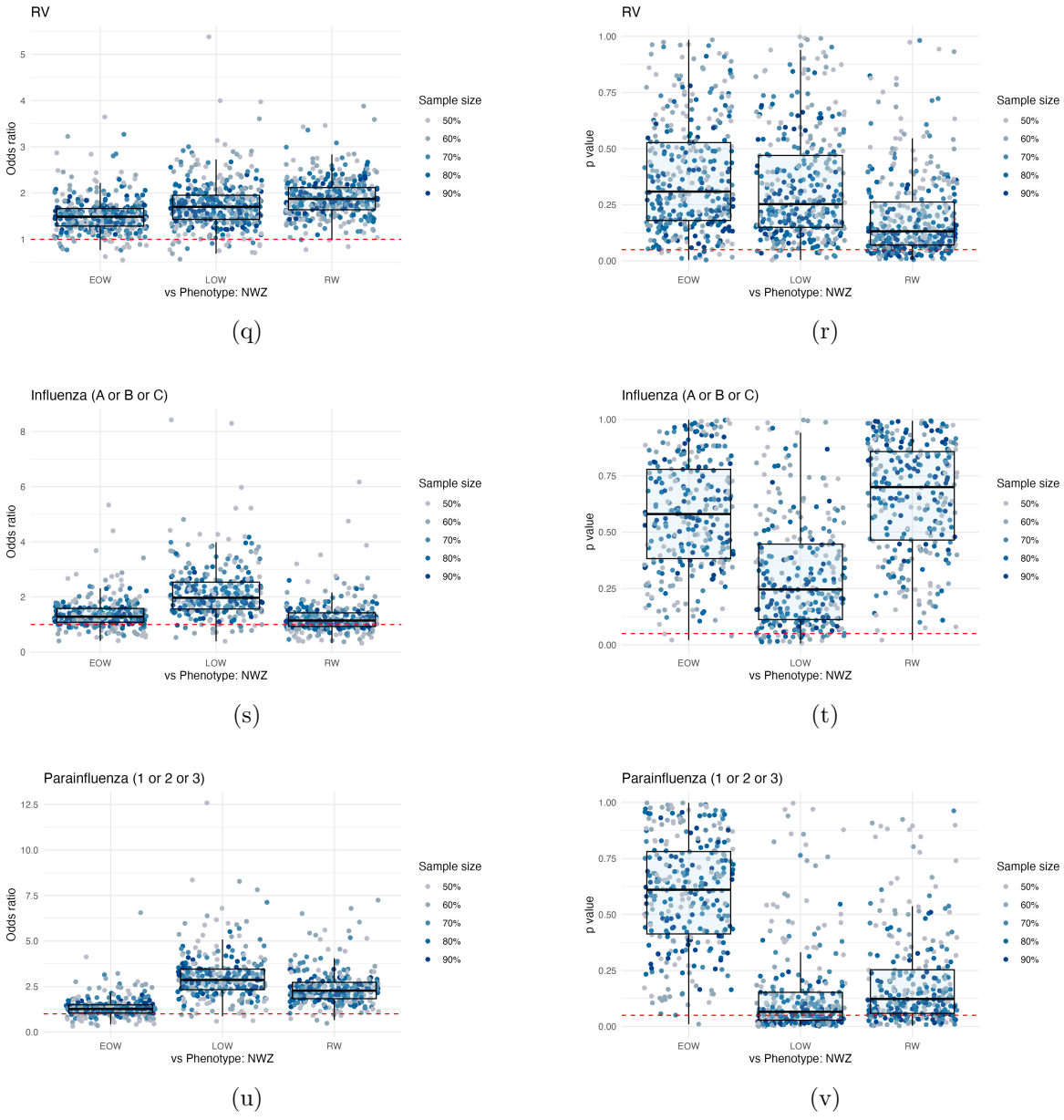
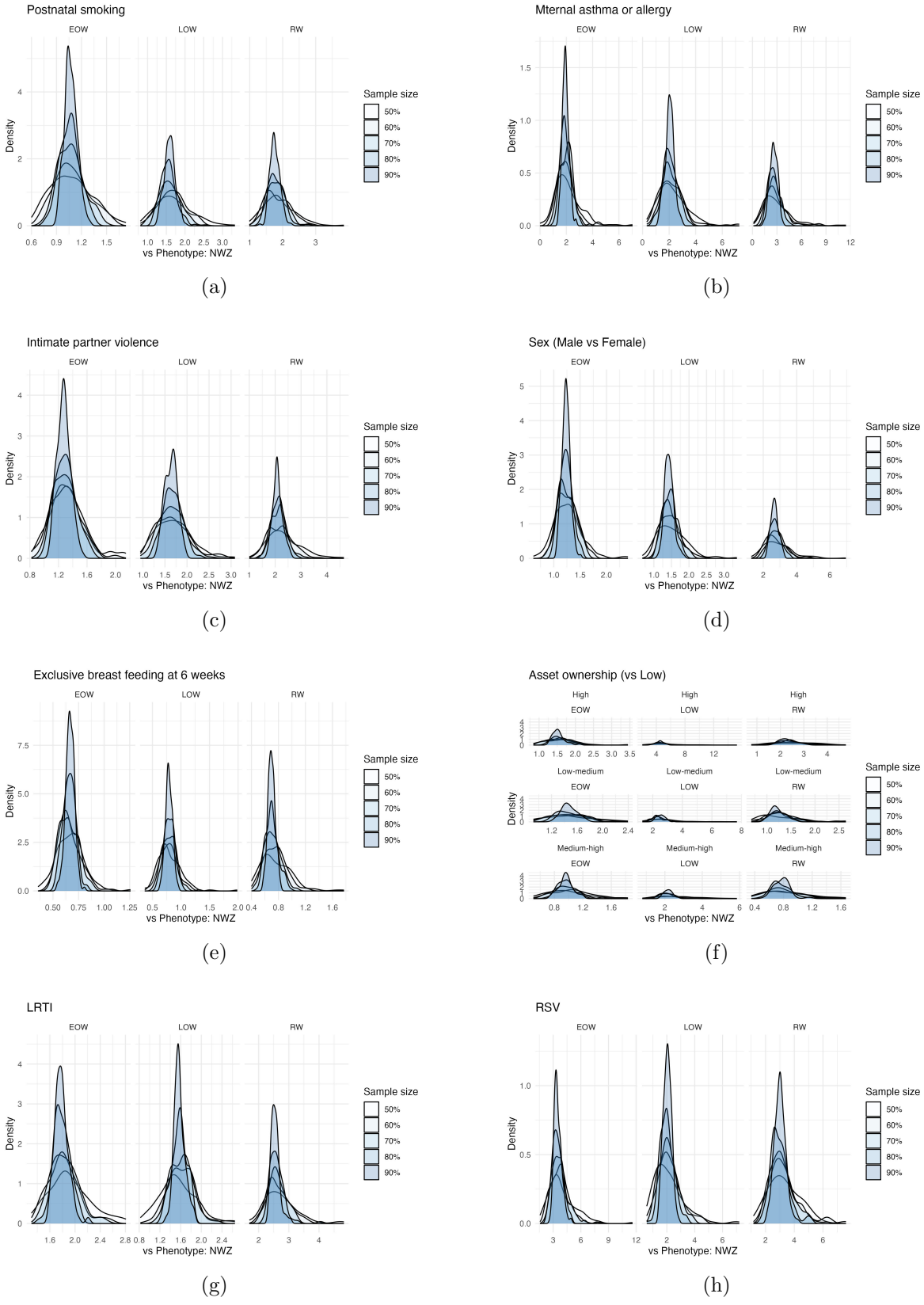
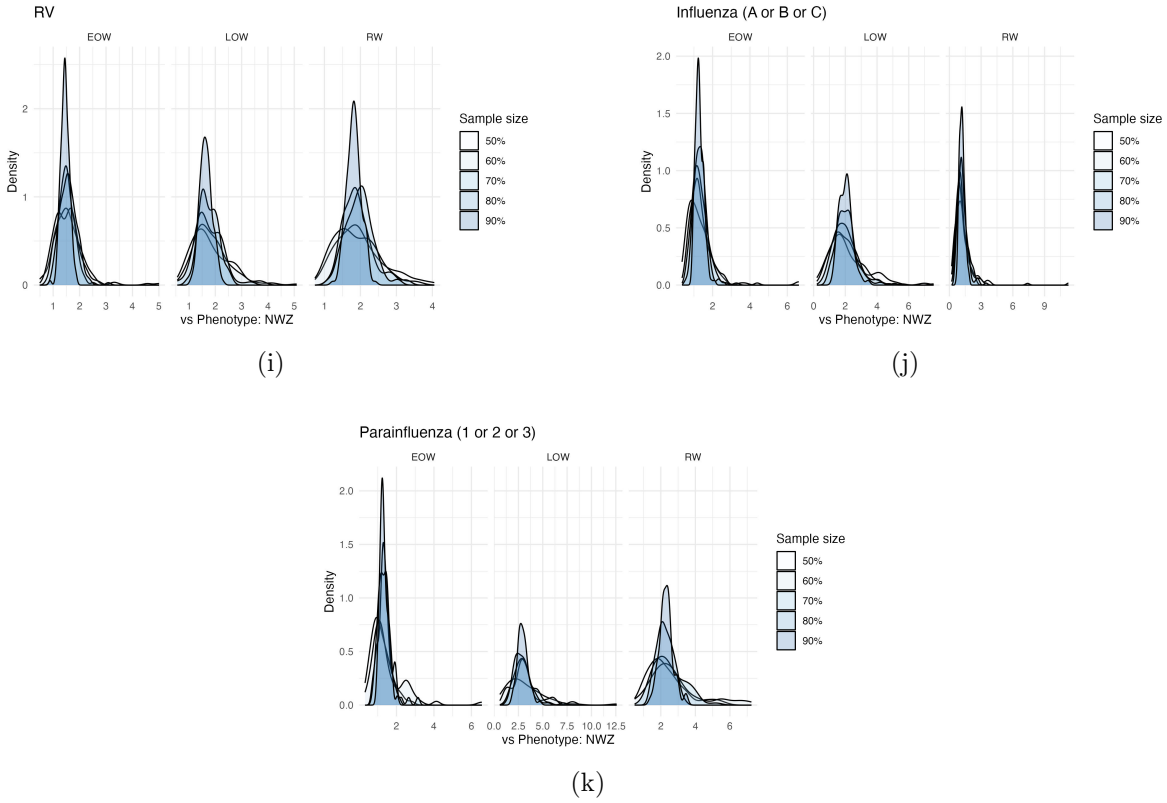


Figure 3.22: Stability analysis sex using cross-validation. Odds ratios (left) and p-values (right).



(a) (b) (c) (d) (e) (f) (g) (h)

Figure 3.23: Density plots of odds ratio estimates by sample size.



(i) (j) (k)

Figure 3.23: Density plots of odds ratio estimates by sample size.

3.11 Discusion

This study is believed to be the first birth cohort to describe patterns of wheezing, associated risk factors, and impact on lung function in an LMIC from birth to age 5 years. Wheezing was common in the DCHS, an African birth cohort from a disadvantaged peri-urban community, with nearly half of infants experiencing at least one episode of wheezing before reaching 5 years of age. However, the frequency of wheezing in the DCHS was lower than in the ALSPAC cohort in the UK, and the prevalence of wheezing declined more rapidly in the DCHS.

To identify homogeneous clusters of children with specific patterns of childhood wheezing, referred to as wheezing phenotypes, the PAM algorithm was adapted for longitudinal data using multi-dimensional indicators. This distance-based method was then compared to phenotypes identified through model-based approaches, specifically LCA and random LCA, using binary variables. The differences in input data may partly explain the differences in clustering solutions. Deriving 6 indicators, which more accurately describe the various temporal features of wheezing, enabled the isolation of intermittent wheezing patterns.

The preferred PAM approach identified four wheeze clusters in DCHS: never wheeze (50.5%), early transient (22.6%), late onset (11%), and recurrent wheeze (15.9%), with differences in the age of onset and absence of a persistent phenotype compared with those in ALSPAC.

Throughout this chapter, the term "wheezing phenotypes" describes derived trajectories of wheezing symptoms based on repeated parent-reported outcomes over time and can only be defined retrospectively once wheezing patterns are observed. These classifications are not biological phenotypes in the strict sense but are retrospective groupings based on symptom patterns.

Choosing the method for developing wheeze phenotypes, whether using PAM or LCA, is an important consideration as it leads to clinical interpretability. The interpretation of the clustering output showed that PAM produced more homogeneous and clinically interpretable phenotypes than LCA. Figure 3.9 demonstrates the lack of clinical interpretability under the LCA approach within the DCHS compared to the chosen solution under PAM in Figure 3.5. Based on the 4-class LCA solution shown in Figure 3.9, the classes lack a clear structure that can be meaningfully compared to those identified in previous studies. The primary statistical objective of identifying phenotypes is to maximise within-class homogeneity. Latent class methods use posterior probabilities to allocate individuals to respective classes. An individual's class membership is therefore not fixed, but based on probability. Instead, each individual is assigned a non-zero probability of belonging to each class. Therefore, it is common practice to assign individuals to the latent class for which they have the highest posterior probability. Therefore, because LCA assigns individuals to classes based on posterior probabilities, there is inherent uncertainty in these class assignments. Some subjects may have low posterior probabilities for all classes, indicating that their wheezing patterns do not clearly align with any specific phenotype. After individuals have been classified into a specific latent class based on the highest posterior probability, their representativeness within the class is considered

equal, even though there may be significant variations in individual posterior probabilities. This suggests that the latent classes might not accurately represent homogeneous wheeze profiles over time. Therefore, it is likely that the derived indicators used in the PAM approach primarily served as precursors for identifying more internally homogeneous phenotypes.

Comparing the results of PAM and LCA showed that PAM produced four clinically interpretable and homogeneous phenotypes, unlike the LCA clustering of binary variables. Importantly, PAM successfully identified a pure group of children who never wheezed, as well as early and late wheezing classes consistent with previous studies on wheezing phenotypes. Additionally, PAM isolated intermittent wheezing patterns into a distinct class, which had previously contributed to significant within-class heterogeneity in the LCA binary data approach.

Although the wheezing phenotypes within the DCHS share the nomenclature of wheezing classes identified in other studies, care should be taken when comparing these wheezing phenotypes to those from other studies which share the same labels because of the uniqueness of the DCHS cohort. The DCHS cohort may have distinct demographic and environmental characteristics that influence the manifestation and progression of wheezing phenotypes. Additionally, differences in study design, data collection methods, and analytical techniques can lead to variations in the identification of wheezing classes. Therefore, while the labels may be similar, the underlying factors and the context in which these phenotypes are observed can differ significantly, necessitating cautious and nuanced comparisons.

A major strength of the study is the derivation of internally homogeneous phenotypes, with the comparison to ALSPAC enabling a direct comparison of the internal structure of the phenotypes by visualising inter-individual wheeze patterns. No such comparison exists in previous LCA studies despite phenotypes sharing similar nomenclature.

Early-life LRTI was one of the strongest associates of all three wheezing clusters, with the strongest association with the recurrent phenotype. RSV-LRTI was associated with early, late, and recurrent wheezing. A previous study found that early-life RSV-LRTI was associated with recurrent wheezing at age 2 years in the DCHS cohort, [24]. The analysis in this chapter extends these findings, showing that this association persists at age 5. By contrast, no other respiratory viruses were associated with early transient, late onset, or recurrent wheezing phenotypes. In cohorts from high-income countries, early-life rhinovirus was associated with ensuing recurrent wheezing or asthma, [130, 131]. However, this association was not observed in DCHS, spotlighting another distinction between LMIC and HIC contexts.

Tobacco smoke exposure has been associated with early transient wheeze and other wheeze phenotypes in HICs, [132]. In this study, maternal smoking prevalence was high, with nearly a third of women self-reporting smoking. This exposure to tobacco smoke emerged as a significant risk factor for late and recurrent wheeze phenotypes. This highlights the imperative to enhance and direct smoking cessation programs toward women of childbearing age. Intimate partner violence was

identified as a risk factor for the recurrent wheeze phenotype, aligning with previous observations in low-income families in the USA that link maternal perceived stress to childhood wheezing, [133]. This also extends previous findings from the DCHS, in which postnatal intimate partner violence was found to be associated with wheezing or recurrent wheezing from birth to age 2 years,[134] and with impaired lung function at 1 year, [135].

A key contribution was the development of a CV framework. The CV framework developed in this chapter proved useful for variable selection and evaluating the stability of model estimates related to wheezing phenotypes. The framework identified and confirmed key risk factors such as LRTI and RSV-LRTI. The analysis demonstrated that while associations like those between male sex and recurrent wheezing remained robust across various subsamples, other associations exhibited variability, thereby providing a nuanced understanding of the reliability of the identified risk factors under different data conditions. This method strengthens confidence in key predictors by ensuring that associations are not the result of specific data samples, but are consistently observed across resampling iterations.

Lung function testing indicated that both early and recurrent wheezing was associated with elevated airway resistance at 5 years, with the most pronounced impact observed in children experiencing recurrent wheezing. Elevated resistance reflects narrowed airways, likely originating from impairments developed early in life, possibly attributable to factors such as LRTI or exposure to smoke. Consistent with this, previous research demonstrated that LRTI in early childhood within the DCHS cohort led to impaired lung function from birth to age 2 years, [24, 65].

A limitation of this study was the use of caregiver-reported wheezing, which might have resulted in over-reporting or under-reporting of wheezing. However, it has previously shown that a strong agreement exists between caregiver-reported wheezing and health worker-ascertained wheezing (κ=0.8) in the DCHS cohort, [24]. The follow-up period was up to 5 years; longer-term follow-up would provide insight into the persistence of trajectories through adolescence. However, the analyses enabled the definition of pre-school wheezing phenotypes over a period when the incidence of LRTI and onset of wheezing are at their highest.

Despite including many potential predictors in a full model to control confounding, the possibility of unmeasured confounding cannot be ruled out. In this analysis, confounding refers to the confounding of predictors in their associations with the outcome, as there is no single exposure being analysed. The effect of removing a variable during the model-building process on the coefficients of the adjusted model was evaluated when possible confounding was suspected, and no discernible impact was observed. Additionally, a DAG was included to illustrate the association between risk factors and wheeze phenotypes. However, it is important to acknowledge that DAGs are based on theoretical assumptions and existing knowledge, which may not fully encompass all relevant factors.

Participants who were excluded exhibited higher socioeconomic status, a factor linked to a greater prevalence of wheezing, [136]. Additionally, this group had lower rates of maternal smoking com-

pared to those included in the study. However, given the high retention rates in the cohort, the fact that the vast majority of participants had complete and comprehensive data collected across 14 time points, and because relocation was the most common reason for loss to follow-up, it is unlikely that selection bias is a major concern.

The findings might not be generalisable to other geographical areas; however, the cohort has many features that are common in other LMICs, including low socioeconomic status, a high burden of infectious diseases, high rates of maternal HIV, psychosocial stressors, and exposure to tobacco smoke. These characteristics are prevalent in many parts of Sub-Saharan Africa, suggesting that the results may be relevant to similar peri-urban, resource-limited settings within the region. It should also be noted that while the DCHS is representative of some such contexts, generalisability to more urbanised or wealthier Sub-Saharan African communities, or areas with higher mobility or different environmental exposures, may be more limited. Further studies are needed to validate these phenotypes and associated risk factors across diverse Sub-Saharan African populations. Despite its limitations, the study benefited from various strengths, including thorough longitudinal phenotyping, high retention rates, attentive surveillance for LRTI with detailed cause delineation, and lung function measurements starting at 6 weeks of age.

The statistical advancements include a CV framework for assessing variable selection stability, and the application and comparison of the statistical strengths and weaknesses of the PAM and LCA clustering approaches in the DCHS. The repeated sampling approach for phenotype classification further refines the assessment of temporal stability in wheezing phenotypes, reducing uncertainty and improving reproducibility. Together, these statistical methods improve the stability, reproducibility, and transparency of phenotype classifications, reducing uncertainty in inference and enhancing the robustness of respiratory epidemiology findings.

The findings in this chapter are novel as (1) the DCHS is the first study of wheezing phenotypes from birth through early childhood in an LMIC where exposures are quite different to those in high-income countries, (2) data was collected at 14 time points in the first 5 years of life, with a high cohort retention rate; (3) there are specific differences between the wheeze phenotypes described and those from high-income cohorts such as ALSPAC, with differences in the age of onset of specific phenotypes, and different viral exposures (with RSV strongly associated with recurrent wheezing, an important finding in the context of new RSV preventive interventions), (4) It is also the only birth cohort in an LMIC in which comprehensive early life lung function was measured preceding wheezing or LRTI, enabling accurate delineation of the impact of these on lung function. The finding that a recurrent wheezing phenotype is associated with impaired lung function at 5 years, independent of baseline lung function, is novel.

Chapter 4

Distinctive lung function trajectories from birth to 5-years in a South African birth cohort study

4.1 Introduction

4.1.1 Chapter aims

This chapter aims to describe the development trajectories of early-life lung function outcomes. Beginning with an examination of the observed lung function outcomes and how their trajectories track, the discussion extends to the strategies employed to address the impact of age-related methodological changes on lung function trajectories over time. Specifically, acknowledging the observed mechanistic shift at 3 years, where measures obtained during quiet sleep before the age of 3 are contrasted with those collected while children are awake and in a seated position after the age of 3.

In addition to age, the trajectories of our outcomes are influenced by several anthropometric factors such as height and weight. FRC (litres), for example, exhibits age-related increments during childhood and is further influenced by height as an indirect indicator of chest size. Due to the strong correlation between age and height, it will be useful to exclude the height component from the developmental trajectories. This decision will facilitate modelling efforts by simplifying the analysis, allowing a concentration solely on age as the primary determinant of lung function outcomes. To illustrate this anthropometric influence, Figure 4.1 presents a three-dimensional illustrative depiction of the relationship between age, height, and lung function outcome, employing FRC (left) and LCI (right) as exemplars. The top left panel, for both plots, illustrates the relationship between age and each lung function outcome, while the bottom right panel shows the relationship between height and each lung function outcome, including data dispersion. Lung volume increases proportionally with height, resulting in a correspondingly higher FRC for taller subjects, as shown by the bottom

right panel in Figure 4.1 (left). While LCI is widely believed to be unaffected by variations in body size as illustrated by the presence of random scatter in the bottom right panel (right).

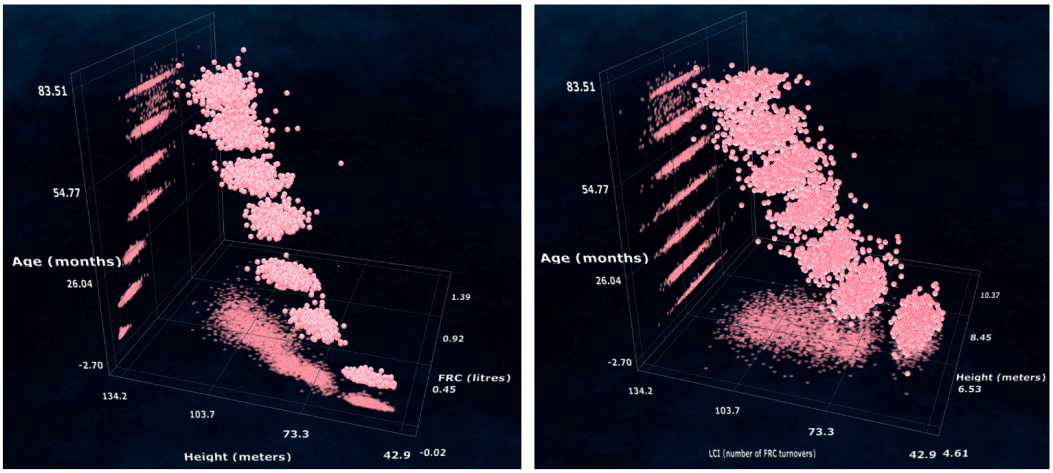


Figure 4.1: Relationship between age, height, and lung function outcomes: FRC (left) and LCI (right) as exemplars.

The initial analysis aims to model the individual development of lung function outcomes from birth to school age using the generalised additive models for location, scale and shape (GAMLSS) and interrupted time series (ITS) approach. To justify the choice of methods, the combination of GAMLSS and ITS is chosen since GAMLSS offers a flexible framework that can accommodate various distributions by modelling the first four moments: location, scale, skewness, and kurtosis. Meanwhile, ITS accounts for the step change observed at three years. It is further hypothesised that the DCHS cohort can be classified into multiple distinct groups of children, each exhibiting notable differences concerning lung function compared to the others. Additionally, it is proposed that each of these distinct groups will possess unique characteristics of their own. This chapter therefore explores the latent trajectories for univariate lung function outcomes. As the analysis is expanded by advancing towards the multivariate framework, it enables the modelling of multiple lung function outcomes simultaneously, while additionally taking into account the correlation between the outcomes.

Lung function measures can be expressed as z-scores, for ease of interpretation, which represent the number of standard deviations from the mean of a healthy reference population. This allows comparisons of a study cohort to a healthy normal population. Before beginning to derive latent trajectories, the aim is to convert individual observations to z-scores. In modelling multiple outcomes, the utilisation of z-scores is useful, since all the outcomes will be placed on a normal distribution curve (i.e. placing them on the same scale). Due to the lack of appropriate reference distributions, z-scores, using the LMS approach, were calculated from the DCHS cohort itself, not from a reference population. In other words, the z-scores show how each subject compares to the other participants

in the DCHS cohort, not to a reference population.

So summarising, the clinical aim is to characterise lung function profiles during the first 5 years of life and examine the underlying heterogeneity. From a statistical perspective, the focus is on methods for standardisation, methods for modelling adjusted changes in profiles due to mechanistic differences in measurement, and methods for identifying heterogeneity.

4.1.2 Details of DCHS lung function testing

A unique aspect of the DCHS is that comprehensive lung function measurements were assessed from 6 weeks of age, with longitudinal measurements done annually, with a very high success rate, Figure 1.1. Measurements included tidal breathing flow volume loops (TBFVL) which measured respiratory rate, tidal volume, and the ratio time of peak total expiratory flow to time of expiration (t_{PTEF}/t_E). Multiple breath washout (MBW) was also conducted to measure functional residual capacity (FRC) and lung clearance index (LCI), an indicator of ventilation homogeneity. Additionally, respiratory impedance (Zrs) was measured using airwave oscillometry (AOS), expressed as respiratory system resistance (Rrs) and reactance (Xrs). Oscillometry included both conventional (multiple-frequency) and intra-breath (single-frequency tracking) methods, [137]. From the multiple-frequency Zrs data, the parameters of resistance and compliance were estimated by model fitting; from intra-breath measurements, end-expiratory resistance (Ree) and reactance (Xee) were determined.

Measurements of infants up to two years old were taken during behaviourally assessed quiet sleep, with the infants lying supine, their heads in a neutral position, and breathing through a face mask and filter. The collection of these measures adhered to current recommended standards for both collection and analysis [138, 139], following previously published protocols with the DCHS, [64]. The preschool measures (3 to 5 years) were collected with children awake and sitting. This is a notable distinction between the assessment methodologies employed in preschool testing and infant testing conducted before the age of two. Tests were conducted and evaluated following the latest recommendations for quality assessment in preschool MBW testing[140] and oscillometry, [141]. Lung function was deferred until at least 4 weeks after an LRTI or wheezing episode.

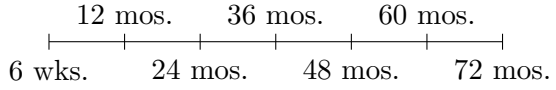


Figure 4.2: Annual lung function assessments.

4.2 Methods

Longitudinal modelling can be approached using various frameworks. Linear mixed effects (LME) models require normally distributed responses, and generalised linear mixed effects (GLME) models extend that to responses that follow distributions from the exponential family, GAMLSS add a whole variety of additional Box-Cox type distributions to enable an exact (or improved) match of distribution to the response and allows for the explicit modelling of moments other than just the mean. GAMLSS methodology is often used to derive reference ranges and z-scores. The GAMLSS framework was selected in this study to model lung function development, as it aligns with the progression towards the intended modelling completion within this chapter. Its flexibility in distribution selection, capacity to model beyond the mean, and established usefulness in lung function research make it suitable for capturing the complexity of longitudinal lung function data.

The GAMLSS framework offers extensive flexibility by accommodating a diverse range of distributions. The focus will be on the Normal, Box-Cox Cole and Green (BCCG), Box-Cox Power exponential (BCPE), and the Box-Cox t (BCT) distributions. The BCCG distribution itself is commonly used in studies involving lung function testing, [142, 143]. The BCCG is founded on the work of Cole and Green, who developed a method to fit a single smoothing term to each of the three distribution parameters, [144].

Each lung function outcome can be standardised to a z-score or transformed into a height-adjusted outcome. A z-score results in a transformed variable that conforms to a standard normal distribution, making it easier to compare and interpret the outcomes. A height-adjusted approach enables the focus to be solely on age as the primary determinant of lung function development, enabling a clearer understanding of the age-related variations. The z-score aids in cross-variable comparison by standardising the data, while the height-adjusted approach allows for a more targeted examination of the age-related aspect of lung function development by removing the confounding effect of height. Using both approaches provides a more comprehensive understanding of the factors influencing lung function outcomes.

Figure 4.3 illustrates two ways lung function outcomes will be adjusted/standardised in this chapter using the FRC outcome as an example. In the subsequent analysis, emphasis will be placed on either the height-adjusted outcome or the z-score, leading to the examination of outcomes on scales distinct from the original lung function measurements. Height-adjusted outcomes will be used when modelling changes in mean lung function development over time, accounting for the effect of a step-change that occurred when testing changed from asleep to awake measurements at age 3 years. Meanwhile, z-scores are used as input into the derivation of latent trajectories. The height adjustment will alter the scale of the outcome, but the overall shape of the trajectory will remain relatively consistent, whilst z-scores reflect deviations from the average trajectory. Although z-scores are useful for assessing deviations from expected lung function, their limitations, particularly regarding interpretability and cross-population comparability, should be acknowledged, [145]. In this thesis, z-scores are used to capture patterns of deviation from a normative trajectory, not to

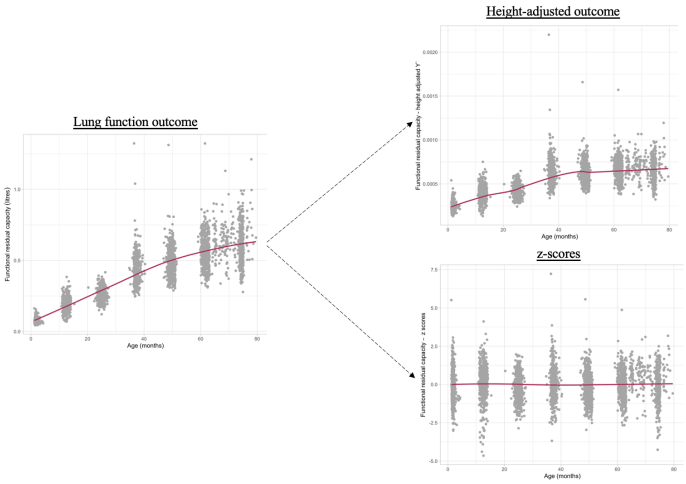


Figure 4.3: Standardisation methods for lung function outcomes with FRC as an example.

draw inferences across populations.

In this chapter, a general overview of the statistical methods is presented first, followed by more detailed explanations of their application to the data in subsequent sections.

4.2.1 GAMLSS framework

When deciding on an appropriate model, an important consideration is the choice of the underlying distribution of the response. GAMLSS relaxes the assumption that the response variable should follow the exponential family distribution assumption. Under the GAMLSS framework, the response variable is allowed to follow a more general distribution family, [146]. The distribution of the outcome variable Y is defined by a set of parameters $\theta = (\mu, \sigma, \nu, \tau)$, which defines the first four moments of a distribution, the location, scale, skewness, and kurtosis. Additionally, these parameters may be modelled as functions of explanatory variables and random effect terms, [147].

There are various different distributional forms available, with the normal distribution as the most basic with parameters μ (location) and σ (scale). Additional distributions can contain up to 4 parameters, μ, σ, and two additional shape parameters ν (skewness) and τ (kurtosis). The BCCG distribution is a popular 3-parameter distribution, while the BCPE and BCT distribution are popular 4-parameter distributions.

The GAMLSS framework assumes that the responses y_i, where $i = 1, ..., n$, are independent conditional on θ_i. The probability density function is given by:

$$f(y_i|\theta_i), \tag{4.1}$$

where θ_i is related to the explanatory variables and random effects through a link function $g_k(.)$, for $k = 1, 2, 3, 4$, shown in Equation 4.2

$$g_k(\theta_k) = \eta_k = \boldsymbol{X}_k\beta_k + \sum_{j=1}^{J_k} \boldsymbol{Z}_{jk}\gamma_{jk}, \tag{4.2}$$

where θ_k and η_k are n dimensional vectors.
$\boldsymbol{X}_k$ represents a design matrix of order $n \times J'_k$ with associated fixed effects β_k of $J'_k \times 1$.
$\boldsymbol{Z}_{jk}$ represents a design matrix of order $n \times q_{jk}$
γ_{jk} represents a $q_{jk} \times 1$ dimensional random effect matrix.
J'_k represents the number of covariates in the fixed effects of η_k and J_k the number of random effects in η_k.

A GAMLSS model with four parameters $f(y|\mu, \sigma, \nu, \tau)$ is written in the following form:

$$g_1(\mu) = \eta_1 = \boldsymbol{X}_1\beta_1 + \sum_{j=1}^{J_1} \boldsymbol{Z}_{j1}\gamma_{j1} \tag{4.3}$$

$$g_2(\sigma) = \eta_2 = \boldsymbol{X}_2\beta_2 + \sum_{j=1}^{J_2} \boldsymbol{Z}_{j2}\gamma_{j2} \tag{4.4}$$

$$g_3(\nu) = \eta_3 = \boldsymbol{X}_3\beta_3 + \sum_{j=1}^{J_3} \boldsymbol{Z}_{j3}\gamma_{j3} \tag{4.5}$$

$$g_4(\tau) = \eta_4 = \boldsymbol{X}_4\beta_4 + \sum_{j=1}^{J_4} \boldsymbol{Z}_{j4}\gamma_{j4} \tag{4.6}$$

The Rigby & Stasinopoulos (RS) algorithm is used to estimate β_k and γ_{jk} through an inner and outer cycle, where the outer cycle maximises the penalised likelihood with respect to β_k and γ_{jk}, sequentially for each θ_k. The RS algorithm is a generalisation of the Mean And Dispersion Additive Model (MADAM) algorithm, [148]. The Cole and Green (1992) algorithm is additionally available.

The GAMLSS framework allows the inclusion of additive smoothing terms, where the explanatory variables are referred to as terms in the model. GAMLSS allows modelling μ, σ, ν, and τ as linear and/or non-linear and/or smoothing functions of the explanatory variables. This enables the explanatory variables to influence the predictors (represented as η's) of the specific parameters, consequently affecting the parameters themselves. The relationships between a predictor η and the terms can be linear or non-linear (or a smoother under non-linear). Smoothers are non-parametric methods that let the data dictate the relationship between predictors and explanatory variables. Additive terms refer to the fact that to evaluate the effect of each explanatory variable on the

predictor for each parameter we have to add up their individual effects, [147]. Additivity in a model does not mean there are no interaction effects.

Given the continuous explanatory variable, Age, the terms of interest (in addition to a constant term b_0) in this study would be

- $b_1 Age$ which is a linear additive term of a continuous variable;
- $S_1(Age)$ smoothing additive terms for Age.

A factor variable with two levels could further be included in the study. Additional terms which relate to the inclusion of the factor variable include:

- b_3 if($f = 2$) the main effect of the factor variable f;
- $s_2(Age)$ if($f = 2$) an interaction of f with smoothing terms for Age.

Both univariate and multivariate smoothers are available within the GAMLSS framework, [147]. A univariate smoother, for example, $f(Age)$, includes only one explanatory variable Age. Examples of univariate smoothers include cubic splines and P-splines. Although not a focus of this study, a multivariate smoother is defined by two or more explanatory variables such as $f(Age, Weight)$. Both univariate and multivariate smoothers can be used as additive terms, where univariate smoothers are the main nonlinear effects of explanatory variables on a distribution parameter (μ, σ, ν, and τ) and multivariate smoothers are their nonlinear interaction effects.

Cubic spline functions can be used for univariate smoothing, [147]. Given a set of real numbers on the interval $[a, b]$, which meets the condition $a < t_1 < t_2 < ... < t_n < b$. A function g which is defined on the interval $[a, b]$ is a cubic spline if on each of the intervals $(a, t_1), (t_1, t2), (t2, t3), ..., (t_n.b)$, g is a cubic polynomial and the cubic polynomials fit together at points t_i such that the first and second derivatives of g are continuous at all points in the interval $[a, b]$. Where the points t_i are called knots, [149].

Cubic splines are the solutions to a minimisation problem. The penalised sum of squares is defined in Equation 4.7.[147, 149]

$$Q_2(g) = \sum_{i=1}^{n} w_i \left(y_i - g\left(t_i\right)\right)^2 + \lambda \int_a^b \left\{g''(x)\right\}^2 dx, \tag{4.7}$$

where $g(t)$ is a twice differential function of t in the previously defined interval $[a, b]$, λ is a smoothing parameter, w_i for $i = 1...n$ are prior weights, and g'' is the second derivative of the function.

Link functions

The link function $g_k(.)$ relates the parameters of the distribution (μ, σ, ν, and τ) to the explanatory variables. Each parameter of the distribution can have its own link function, connecting the parameters to the explanatory variables through additive models. If the distribution in question has only one parameter, the model reduces to a standard generalised additive model (GAM), [150]. Link functions transform the parameters of the distribution to ensure they remain within valid ranges.

There are several types of link functions commonly used in GAMLSS: Identity Link (directly relates the parameter to the explanatory variable without transformation), Log Link (ensures the parameter remains positive by modelling the logarithm of the parameter), Logit Link (used for parameters that need to be constrained between 0 and 1). In the context of lung function outcomes which generally represent positive continuous variables, the log link function is a suitable choice as it ensures that the predicted values remain positive.

4.2.2 The LMS method

The LMS method is a unique modelling framework within the GAMLSS framework where the distribution of the outcome variable of interest is defined using three time-varying parameters (ν, μ, and σ). Kurtosis is not modelled under this approach. Within GAMLSS, the LMS method is equivalent to assuming the Box-Cox Cole and Green distribution (BCCG) for the response variable. The LMS method can be used for the standardisation of anthropometric measures where μ, σ, and ν are from a reference population. In this study, this method is transferred to model lung function development. The following formula applies a transformation to the response variable Y, resulting in the transformed variable Z that conforms to a standard normal distribution.

$$\begin{aligned} Z &= \frac{1}{\sigma v}\left[\left(\frac{y}{\mu}\right)^{v} - 1\right] \quad && \text{for all } v \neq 0 \\ Z &= \frac{1}{\sigma}\log\left(\frac{y}{\mu}\right) && \text{for all } v = 0 \end{aligned} \tag{4.8}$$

where $y > 0$, $\mu > 0$, $\sigma > 0$, and $-\infty < \nu < \infty$. We now assume that the parameters ν, μ, and σ vary smoothly with age (t), and can therefore be replaced in Equation 4.8 by the smooth curves $L(t)$, $M(t)$ and $S(t)$ as:

$$\begin{aligned} Z &= \frac{[y/M(t)]^{L(t)} - 1}{L(t)S(t)} \quad && \text{for all } L(t) \neq 0 \\ Z &= \frac{\log[y/M(t)]}{S(t)} && \text{for all } L(t) = 0 \end{aligned} \tag{4.9}$$

Equation 4.9 transforms y into a skewness-corrected standardised value at each time point. For a set of n independent measurements y_i at corresponding covariate values t_i, the log-likelihood function from Cole and Green (1992) can be derived from Equation 4.9 and is expressed as follows [144]:

$$l = l(L, M, S) = \sum_{i=1}^{n} \left(L\left(t_i\right) \log \frac{y_i}{M\left(t_i\right)} - \log S\left(t_i\right) - \frac{1}{2} z_i^2 \right) \tag{4.10}$$

Where z_i denotes the standard deviation scores corresponding to y_i. The penalised likelihood is given by

$$l - \frac{1}{2}\alpha_\lambda \int \left\{L''(t)\right\}^2 dt - \frac{1}{2}\alpha_\mu \int \left\{M''(t)\right\}^2 dt - \frac{1}{2}\alpha_\sigma \int \{S(t)\}^2 dt \tag{4.11}$$

The $L(t)$, $M(t)$ and $S(t)$ curves are estimated using maximum penalised likelihood. Where α_λ, α_μ, and α_σ are smoothing parameters.

4.2.3 The Box-Cox T Method

The Box-Cox t (BCT) distribution is an extension of the Box-Cox normal distribution, [151, 144]. It serves as a versatile model for capturing both skewness and kurtosis. The BCT distribution is defined for $Y > 0$ and characterised by its four parameters BCT(μ, σ, ν, τ).

A positive random variable Y with a BCT distribution may be represented through the transformed random variable Z, as shown in Equation 4.12 below.

$$\begin{aligned} Z &= \frac{1}{\sigma v}\left[\left(\frac{Y}{\mu}\right)^v - 1\right], \quad \text{if } v \neq 0 \\ Z &= \frac{1}{\sigma}\log\left(\frac{Y}{\mu}\right), \qquad \text{if } v = 0. \end{aligned} \tag{4.12}$$

For $0 < Y < \infty$, where $\mu > 0$ and $\sigma > 0$, the random variable Z follows a truncated t-distribution with τ degrees of freedom. The distribution of the transformed variable Z is a key difference between the BCT and BCCG distributions, here Z follows a truncated t-distribution as opposed to a standard normal distribution under BCCG. Here, $\tau > 0$ is treated as a continuous parameter. The probability density function of $Y \sim BCT(\mu, \sigma, \nu, \tau)$ is given by

$$f_Y(y) = f_Z(z)\left|\frac{dz}{dy}\right| = \frac{y^{v-1}}{\mu^v \sigma} f_Z(z) \ , \tag{4.13}$$

where $f_Z(z)$ is the truncated t probability density function of the transformed variable Z given by Equation 4.14.

$$f_Z(z) = \frac{f_T(z)}{F_T(1/\sigma|v|)} \ , \tag{4.14}$$

for $-\frac{1}{\sigma\nu} < Z < \infty$ if $\nu > 0$,
and $-\infty < Z < -\frac{1}{\sigma\nu}$ if $\nu < 0$, where $\sigma > 0$
and $-\infty < \nu < \infty$.

4.2.4 Model selection

GAMLSS models can be compared using the generalised Akaike information criterion $(GAIC) = -2\hat{\ell} + (k, df)$, where $\hat{\ell} = \sum_{i=1}^{n} \log f\left(y_i \mid \hat{\mu}_i, \hat{\sigma}_i, \hat{\nu}_i, \hat{\tau}_i\right)$ is the fitted log-likelihood function and k is a required penalty. If models are nested, they can be compared using the global deviance $(GD) = -2\hat{\ell}$ instead. The model with the smallest GAIC is the preferred choice.

The standard Akaike Information Criterion (AIC) is applied when $k = 2$ and the Bayesian Information Criterion (BIC) is used when $k = log(n)$. Statistical software packages generally apply a penalty of $k = 2$ as default. The choice of k influences the balance between model fit and complexity.

4.2.5 Interrupted time series (ITS) regression

Background

ITS is an approach widely applied in medical research [152, 153], particularly as a quasi-experimental approach for evaluating the longitudinal effects of interventions. In this chapter, the ITS approach is applied in a slightly different manner from the traditional community intervention setup, since there are not multiple communities or subgroups of interest that are repeatedly assessed, with the intervention being introduced to only one subgroup at a time. The counterfactual, modelled by extrapolating the pre-intervention trend, is therefore not of great interest in this analysis.

Describing the effect of the interruption

Effectively, ITS is a pre- and post-comparison method, illustrated by Figure 4.4. This figure visually represents the impact of an interruption on the t_{PTEF}/t_E by comparing the trends and levels before and after the interruption point. The form of the level change at the point of interruption is described by *permanence* and *immediacy*. Permanence refers to whether the effect of the interruption is permanent or temporary. Here, the mechanistic shift at 3 years is assumed to be a permanent change in the level. Immediacy refers to whether the change at the interruption point has an immediate effect or if there is a lagged effect. Here, the level change is known to be an immediate effect.

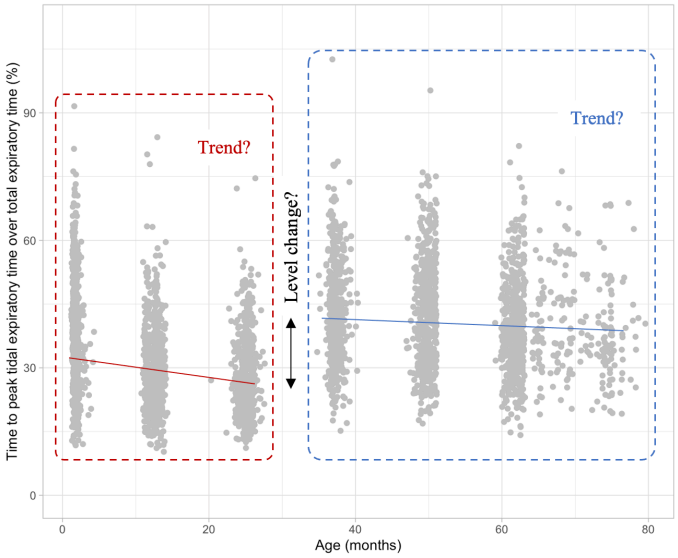


Figure 4.4: Illustration of the interrupted time series design

Methods

Investigating early life trajectories involved utilising the versatility offered by the GAMLSS approach [146] in combination with an interrupted time series (ITS) approach, [153]. This facilitated the modelling changes in age-related methodologies on lung function trajectories over time. In a traditional ITS study, an outcome of interest (a continuous lung function measure in this thesis) is used to establish an underlying trend, which is "interrupted" at a known point in time. Under this approach, accounting for the impact of a 'step change' that occurred at 3 years was possible, coinciding with the shift from asleep infant measures to awake preschool measures during testing. ITS allows for different slopes before and after the 3-year time point.

At least three variables are required:

- Age: the time elapsed since the start
- X_t : a dummy variable indicating the different periods (pre-interruption is coded using 0 and post-interruption is coded using 1).
- Y_t : the outcome of interest at time t.

The following interrupted regression model is used in a simple analysis:

$$Y_t = \beta_0 + \beta_1 Age + \beta_2 X_t + \beta_3 Age X_t \ , \tag{4.15}$$

where β_0 represents the baseline level at Age = 0, β_1 represents the change in outcome associated with a time unit increase (pre-interruption trend), β_2 represents the level change around the interruption point, and β_3 represents the slope change following the interruption point (through the interaction $AgeX_t$).

Figure 4.4 illustrates the level change of interest under the ITS approach. Exactly where this level change occurs is specified manually by the statistician/clinician. For example, if suspecting this change to occur after 3 years, specifying this point in the model can be done through a dummy variable.

4.2.6 Impact of height on lung function

There are several lung function measurements which depend on both body size and age and are standard considerations in longitudinal lung growth trajectories. For the lung function outcomes of interest in this study, height was used as a proxy for body size.

To account for the impact of height on lung function, the adjustment of lung function measurements involved removing the height component by extending a multiplicative model approach described by Cole et al. [70], where the fitted model is described in Equation 4.16. The multiplicative model approach allowed for isolating the specific influence on lung function that is not solely attributable to variations in height, which is an established predictor of certain lung function measures.

Given Y as the lung function outcome for an individual at a given age, a model is constructed for the $E(Y)$ as

$$\begin{aligned} \mu &= e^{\alpha_\mu + a \cdot I} Height^b e^{cs(age)(c+d \cdot I)}, \text{ or equivalently,} \\ \mu &= e^{\alpha_\mu} e^{a \cdot I} Height^b e^{c \cdot cs(age)} e^{d \cdot I \cdot cs(age)} \\ log(\mu) &= \alpha_\mu + I \cdot \alpha + b \cdot \log(Height) + c \cdot cs(\log(age)) + d \cdot I \cdot cs(log(age)), \end{aligned} \tag{4.16}$$

where $log(\mu) = log(E(y))$, I is a dummy variable indicating the different time periods on either side of age 3 and $cs()$ represents non-parametric cubic spline functions of age. A child-specific random effect was included to account for individual deviations in lung function trajectories.

Utilising the model introduced in Equation 4.16, the division of lung function by $Height^b$ can be performed, where b is the estimated coefficient for $\log(Height)$ from the model for $log(\mu)$. By performing this action, a new variable Y' (shown by Equation 4.17) is created, which is modelled as a function of age.

$$Y' = \frac{Y}{Height^b} \ , \tag{4.17}$$

where Y' represents the height-adjusted lung function variable we want to model. Equation 4.17 is expected to adjust the scale of each outcome, while maintaining the overall shape of the original trajectory. Although it is acknowledged that height alone may not capture the entire spectrum of body size variations, an important consideration was not to eliminate all variation between the longitudinal profiles by standardising all the differences away.

4.2.7 Latent class mixed models (LCMM)

LCMM provides a model-based framework to identify subgroups of longitudinal trajectories, each with its own mean profiles, [154]. The LCMM package, developed by Proust-Lima et al. (2015), offers a range of functions for estimating statistical models based on linear mixed model theory. This approach, often termed a growth mixture model (GMM) among other names, is part of a larger category of longitudinal growth models, [155]. This chapter will utilise this package extensively. LCMM offers a framework to model a range of longitudinal outcomes, including both Gaussian and non-Gaussian quantitative data, as well as ordinal data. In modelling the trajectory, LCMM can accommodate higher-order polynomials, splines, or the beta cumulative distribution function. Random effects are handled with a default unstructured variance-covariance matrix which may vary across latent classes, although a diagonal matrix can be specified. Correlation between errors can also be modelled.

The LCMM approach makes use of three different underlying models. The models are introduced following the notation used by Proust-Lima et al. (2015), [154]. The first model is a structural model, represented in Equation 4.18, which accounts for repeated observations within individuals. Here $\Lambda_i(t)$ denotes the latent process.

$$\Lambda_i(t) = X_{Li}(t)^\top \beta + Z_i(t)^\top u_i + w_i(t) \quad , \quad \forall t \in R \tag{4.18}$$

In Equation 4.18, the vectors of covariates $X_{Li}(t)^\top$ and $Z_i(t)^\top$ are associated with the vector of fixed effects β and the vector of random effects u_i respectively.

The second model is a measurement model, shown in Equation 4.19, which is the link between the latent process and the observed value Y_{ij} at the measurement time t_{ij}.

$$Y_{ij} = H\left(\Lambda_i\left(t_{ij}\right) + \epsilon_{ij}; \eta\right), \tag{4.19}$$

where H is a parameterised link function. Several different link functions (linear, beta, splines, thresholds) can be chosen depending on the type of each longitudinal outcome. When the marker for which you have longitudinal data is continuous, H^{-1} is a monotonic increasing continuous function among either the linear transformation which reduces to the linear mixed model, the rescaled cumulative distribution function (CDF) of a Beta distribution, or the basis of quadratic I-splines with m knots, [154]. Within the LCMM framework, we assume a parameterised link function for linking the outcome with the underlying latent process which it measures. Where ϵ_{ij} are independent normally distributed errors.

The third model, a multinomial logistic model, describes latent class membership as shown in Equation 4.20.

$$\pi_{ig} = P\left(c_i = g \mid X_{ci}\right) = \frac{e^{\xi_{0g} + X_{ci}^T \xi_{1g}}}{\sum_{l=1}^{G} e^{\xi_{0l} + X_{ci}^T \xi_{1l}}} \tag{4.20}$$

The study includes N subjects $(i, i = 1, ..., N)$, with the assumption that the population is composed of G underlying latent classes $(g, g = g, ..., G)$ which are characterised by G mean profiles of developmental trajectories. A discrete latent variable c_i for the latent group structure, where $c_i = g$ if subject i belongs to latent class g. A key distinction from a typical linear mixed model is that in this case, both the fixed effects and the distribution of the random effects are class-specific.

The LCMM framework has been extended to accommodate multiple outcomes that measure the same latent process, [154]. It is assumed that the latent process $\Lambda_i(t)$, defined above, generated K longitudinal outcomes instead of a unique one.

In the multivariate setting, the structural model in Equation 4.18 is identical. However, the measurement model from Equation 4.19 needs to be extended for the multivariate setting, to consider the relationship between the latent process and each longitudinal outcome. Let each subject i $(i = 1, ..., N)$, of outcome k $(k = 1, ..., K)$, at occasion j $(j = 1, ..., n_{ki})$ be represented by Y_{kij}. The age at which the measurements are collected is represented by t_{kij}. This model specification accommodates multiple measurements and associated time records that may vary among subjects and outcomes. The measurement model can be written as follows:

$$\tilde{Y}_{kij} = \Lambda_i \left(t_{kij}\right) + X_{Yi} \left(t_{kij}\right)^\top \gamma_k + b_{ki} + \epsilon_{kij} \ , \tag{4.21}$$

where $X_{Yi} \left(t_{kij}\right)$ are covariates and γ_k is referred to as *contrasts* (which are outcome specific) considering the varying impacts of covariates and time on outcomes, following adjustment for the latent process level. The random effects are represented by $b_{ki} \sim \mathcal{N}\left(0, \sigma^2_{\alpha_k}\right)$ and $\epsilon_{kij} \sim \mathcal{N}\left(0, \sigma^2_{\epsilon_k}\right)$ represents the measurement error.

An outcome-specific parameterised link function is represented by H_k:

$$Y_{kij} = H_k \left(\tilde{Y}_{kij}; \eta_k\right) \ , \tag{4.22}$$

where H_k can be defined using linear, beta, or spline link functions. This flexible specification enables H_k to model a variety of outcome types across different scales. The parameters of the non-linear link functions and the latent process are estimated simultaneously through maximum likelihood (ML) estimation. This approach can also accommodate non-linear fixed effects by utilising splines and a beta link function.

Model selection criteria

In studies employing LCMM to identify lung function trajectories, the BIC is commonly used for model selection, [73, 156]. In this study, additional criteria will be applied. The optimal model will be identified by simultaneously considering:

1. Likelihood-based criteria (such as the AIC and BIC) and model entropy,
2. Minimum class sizes, and

3. Clinical relevance (e.g., if two models yield comparable fit indices, which model would offer the greatest clinical value?) and interpretation (e.g., a model with too many latent classes might fit the data well but be difficult to interpret meaningfully).

This study utilises a minimum class size threshold set at 3%. While this figure may be somewhat arbitrary, the choice is based on considering class sizes below this threshold to be overly speculative from a clinical standpoint. As an example, 3% of a sample size of 950 children is approximately 29 children.

4.3 Results

4.3.1 Understanding developmental trajectories of childhood lung function

Background

The analysis presented in this section focuses on describing the effect of early-life risk factors on lung function development from birth to school age using the GAMLSS and ITS approach. This section of the chapter provides a more detailed discussion of the developmental trajectories and risk factor analysis that was completed and reported by McCready et al. (2024), [157].

Figure 1.1 illustrates the cohort flow for lung function assessments. Retention and testing success rates were high, with 923 (81%) children having $\geq$2 lung function measurements between birth and 5 years. This analysis presented in this chapter included 966 (84.5%) children who had at least one lung function measurement. The objective was to minimise the exclusion of children as much as possible. The single measurements do contribute to the estimation of mean levels at a given age and under the assumption of random missingness, do not detract from the estimation of time profiles. There were 13384 lung function measurements available: 3983 tidal breathing, 3634 MBW, and 3034 oscillometry measurements, Figure 1.1.

Longitudinal patterns of lung function outcomes

Lung function trajectories in infants can be described by visually inspecting the patterns and changes over time. By analysing measurements taken at various stages, the trajectory can be characterised as stable, improving, declining, or exhibiting catch-up growth. Describing the lung function trajectories involves identifying trends, variations, and potential factors influencing the observed trajectories. Figure 4.5 illustrates the lung function trajectories by age. A shift in measurements can be observed between the 2-year and 3-year marks, as denoted by a dotted line. This dotted line signifies a modification in the testing methodology from asleep infant measures to awake preschool measures during testing. Table 4.1 provides details of lung function at each time point, including the number of participants, the percentage of the cohort, and the mean values with standard deviations.

Similar trends emerge across various outcome groups, Figure 4.5. Specifically, FRC, compliance, and tidal volume exhibit ascending patterns with age, with a slight reduction in the rate of increase as age advances. After three years, there is an increase in the dispersion of measured points. Resistance and Ree exhibit initially high measurements at birth, followed by a decline from birth to three years, after which a more linear trend emerges. The dispersion of measured points stabilises after three years. Respiratory rate follows a similar pattern to resistance and Ree but experiences increased dispersion in measured points after three years. LCI and t_{PTEF}/t_E maintain consistent, stable trends across age, with a noticeable jump in the t_{PTEF}/t_E measurements after three years. The dispersion of measured points remains stable for both LCI and t_{PTEF}/t_E. Xee stands out as a unique outcome due to its negative value recordings. Initially, Xee data display relatively high dispersion, which diminishes and becomes more stable after three years.

Table 4.1: Summary of lung function measurements by visit.

Parameter	Age	Frequency	Percentage[1]	Mean (sd)
Functional residual capacity (litres)				
	6-weeks	734	76.0%	0.077 (0.016)
	1-year	593	61.4%	0.20 (0.043)
	2-years	599	62.0%	0.26 (0.045)
	3-years	449	46.5%	0.43 (0.90)
	4-years	557	57.7%	0.48 (0.096)
	5-years	702	72.7%	0.59 (0.13)
Lung clearance index (number of FRC turnovers)				
	6-weeks	734	76.0%	7.15 (0.44)
	1-year	593	61.4%	6.77 (0.56)
	2-years	599	62.0%	6.69 (0.47)
	3-years	449	46.5%	7.00 (0.77)
	4-years	557	57.7%	6.93 (0.59)
	5-years	702	72.7%	6.84 (0.56)
Resistance ($hPa.s.L^{-1}$)				
	6-weeks	677	70.1%	47.59 (14.89)
	2-years	568	58.8%	26.19 (10.25)
	3-years	425	44.0%	11.48 (2.21)
	4-years	614	63.6%	10.13 (2.13)
	5-years	750	77.6%	8.76 (1.93)
Compliance ($L.kPa^{-1}$)				
	6-weeks	677	70.1%	0.001 (0.001)
	2-years	568	58.8%	0.002 (0.001)
	3-years	425	44.0%	0.005 (0.002)
	4-years	614	63.6%	0.006 (0.002)
	5-years	750	77.6%	0.007 (0.002)
t_{PTEF}/t_E (%)				
	6-weeks	781	80.9%	37.96 (12.07)
	1-year	630	65.2%	29.39 (10.09)
	2-years	621	64.3%	27.28 (8.91)
	3-years	494	51.1%	42.84 (11.82)
	4-years	668	69.2%	42.18 (12.13)
	5-years	777	80.4%	38.92 (10.93)
Tidal volume (millilitres)				
	6-weeks	781	80.9%	34.80 (6.34)
	1-year	630	65.2%	92.95 (14.22)
	2-years	621	64.3%	119.46 (16.59)
	3-years	506	52.4%	178.56 (28.97)
	4-years	668	69.2%	202.27 (33.77)
	5-years	777	80.4%	229.43 (49.38)
Respiratory rate (breaths per minute)				
	6-weeks	781	80.9%	48.83 (11.29)
	1-year	630	65.2%	29.51 (5.04)
	2-years	621	64.3%	26.33 (4.32)
	3-years	499	51.7%	28.01 (5.72)
	4-years	667	69.1%	26.66 (6.19)
	5-years	777	80.4%	27.73 (8.00)
Ree ($hPa.s.L^{-1}$)				
	6-weeks	498	51.6%	45.46 (12.59)
	2-years	567	58.7%	19.89 (6.96)
	3-years	394	40.8%	12.32 (2.77)
	4-years	580	60.0%	10.59 (2.54)
	5-years	694	71.8%	8.93 (2.36)
Xee ($hPa.s.L^{-1}$)				
	6-weeks	498	51.6%	-8.21 (8.0)
	2-years	567	58.7%	-0.76 (3.31)
	3-years	394	40.8%	-2.63 (1.71)
	4-years	580	60.0%	-2.08 (1.53)
	5-years	694	71.8%	-1.46 (1.24)

[1]A denominator of 966 active children was used.
t_{PTEF}/t_E = Ratio time of peak total expiratory flow to time of expiration; Ree = respiratory resistance at the end of expiration; Xee = respiratory reactance at the end of expiration; sd = standard deviation

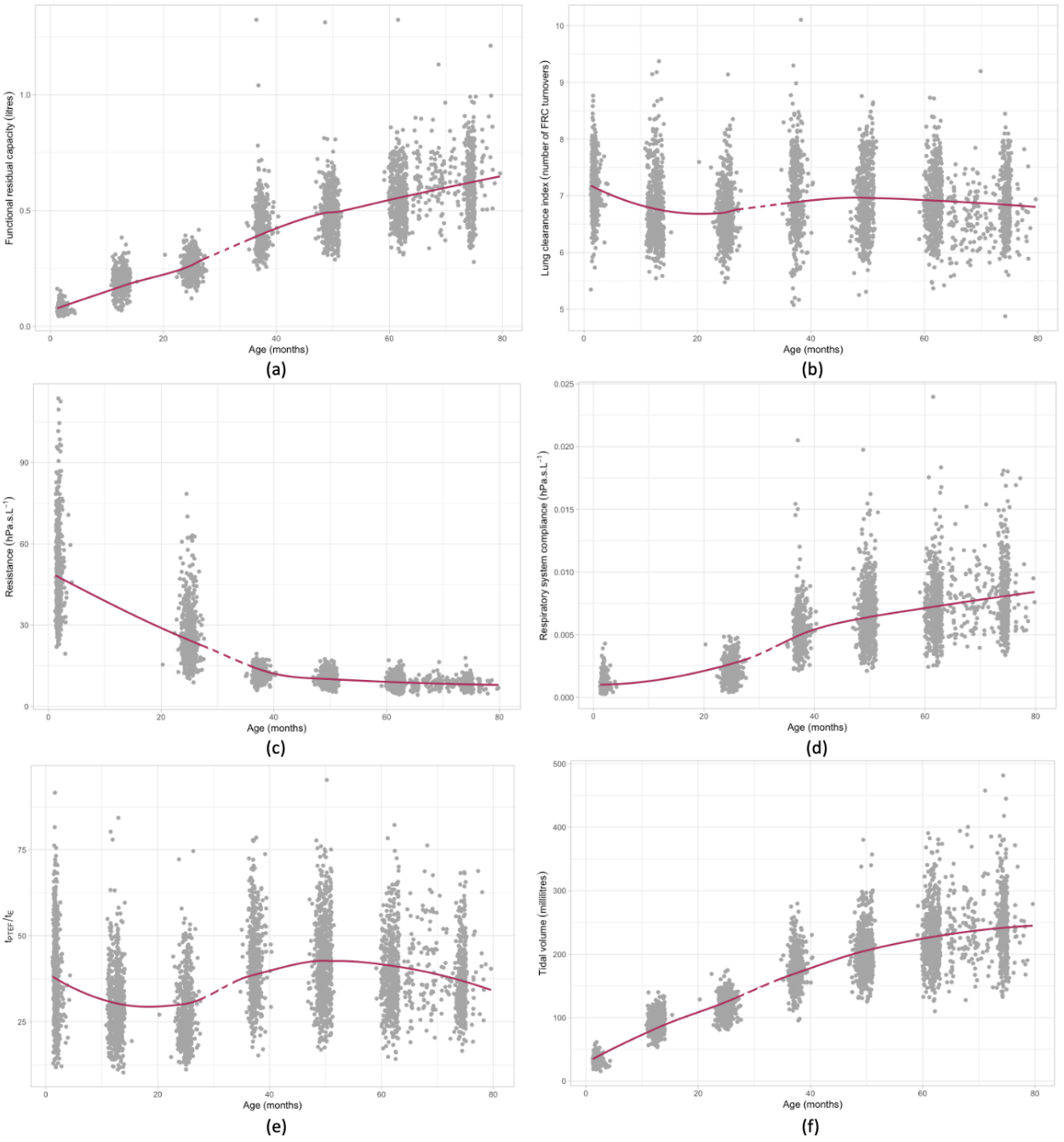


Figure 4.5: Longitudinal plots of measured lung function outcomes (a) functional residual capacity, (b) lung clearance index, (c) resistance, (d) compliance, (e) ratio time of peak total expiratory flow to time of expiration (t_{PTEF}/t_E), and (f) tidal volume versus age. The dotted line between 2 years and 3 years highlights the step change between testing methods.

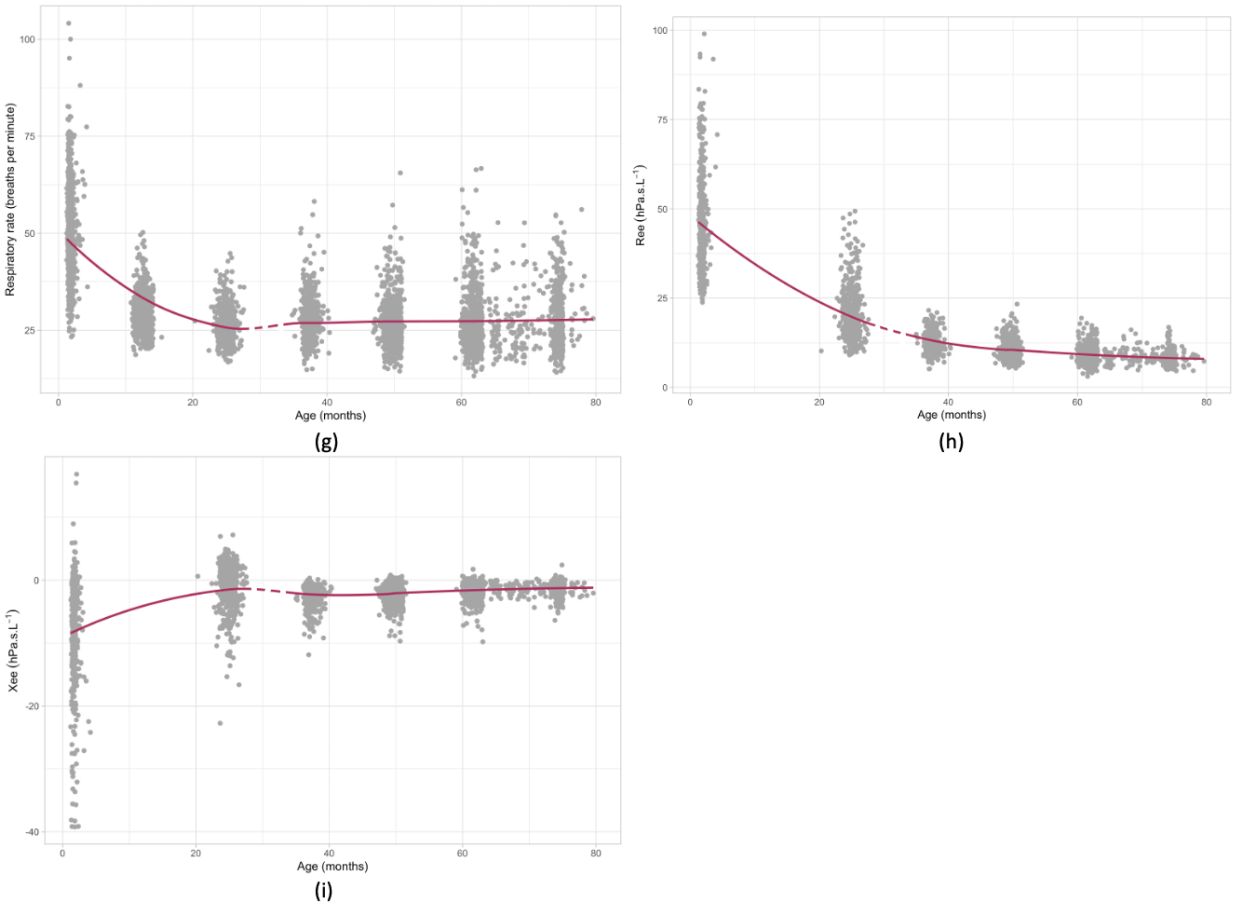


Figure 4.5: Longitudinal plots of measured lung function outcomes (g) respiratory rate, (h) respiratory resistance at the end of expiration (Ree), and (i) respiratory reactance at the end of expiration (Xee) versus age. The dotted line between 2 years and 3 years highlights the step change between testing methods.

4.3.2 Phases in lung function trajectories from birth

Background

In general, lung function follows a three-phase trajectory: 1) growth from birth to early adulthood, 2) a plateau during mid-adulthood, and 3) a gradual decline with ageing, [158]. This study focuses on the growth phase from birth to school age, during which lung function improves as the lungs develop.

In healthy individuals, lung function tends to follow a specific percentile range established early in life, [159]. Essentially, a child's lung performance compared to their peers is largely set in their early years and remains relatively consistent over time. Consequently, a child's starting lung function at birth is therefore believed to be an important determinant of lung function into later childhood. This relative consistency is regarded as the general case; however, it does not preclude the possibility that a child may exhibit catch-up growth over time. The latter part of this chapter, which explores latent classes of lung function, challenges this idea of relative consistency to an extent by seeking to identify smaller groups of children who exhibit either improvement or difficulties over time.

Lung function quartiles

It was investigated how lung function growth may track throughout childhood in the DCHS by calculating quartiles using the 6-week test results. The quartiles which were calculated at 6 weeks were then tracked through early childhood by comparing the mean lung function at each annual visit between the quartiles.

At a fundamental level, exploring quartiles helps determine if infants with initially poor lung health demonstrated improvement or "catch-up" in their lung function during their first year of life and beyond. Figure 4.6 illustrates the trajectories of quartiles calculated at 6 weeks for each lung function outcome. The convergence of the quartiles in Figure 4.6 reflects the narrower spread of measured lung function values in the older awake children compared to the infant measures. When looking at the preschool measures alone the differences in quartiles are clearer. In general, the quartile plots reveal that infants with higher initial values at 6 weeks tend to exhibit higher lung function trajectories as they age, while the opposite holds for infants with lower initial values. This aligns with the hypothesis that starting lung function, alongside LRTI and its pathogenic stratifications, environmental, and maternal risk factors, plays a role in shaping lung function during early childhood. Research on lung function trajectories has revealed two key insights: human lung function follows predictable patterns throughout life, and certain risk factors can influence these patterns, [159]. The middle part of this chapter attempts to identify the factors which influence developmental patterns.

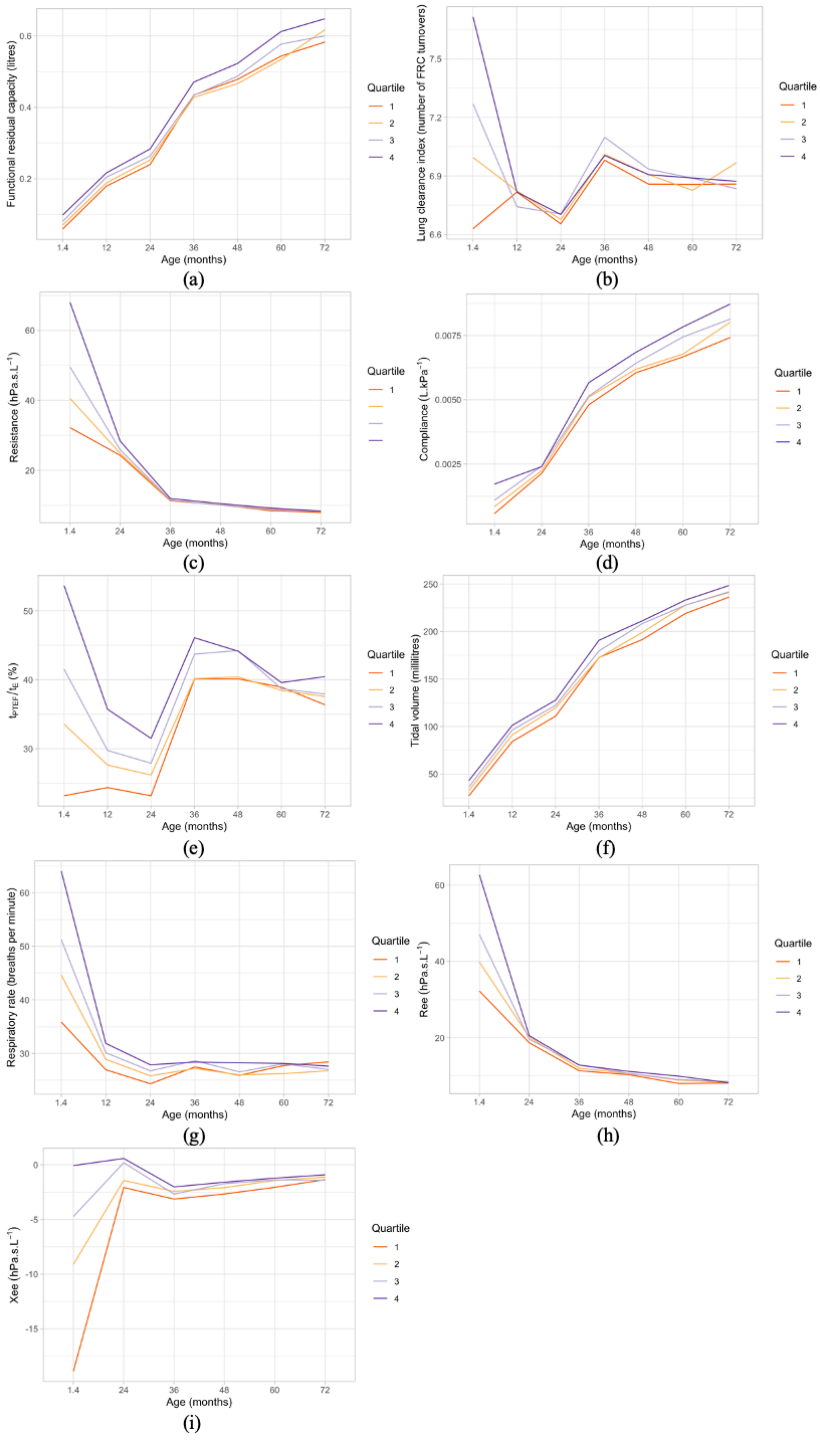


Figure 4.6: Quartile plots of measured lung function outcomes (a) functional residual capacity, (b) lung clearance index, (c) resistance, (d) compliance, (e) ratio time of peak total expiratory flow to time of expiration (t_{PTEF}/t_E), (f) tidal volume, (g) respiratory rate, (h) respiratory resistance at the end of expiration (Ree), and (i) respiratory reactance at the end of expiration (Xee) versus age.

4.3.3 The impact of height on lung function development

Background

Recognising the intertwined nature of height and lung function measurements, this analysis examines the need for height adjustments on various respiratory parameters in children to ensure an accurate interpretation of their lung health. To account for the impact of height (as a proxy for body size) on lung function, the adjustment of lung function measurements involved removing the height component using the multiplicative model, as described earlier by Equation 4.16. This allows for a more precise assessment of respiratory function by isolating the effect of height from other variables influencing lung health.

Results

A child's height is strongly associated with several lung function outcomes such as FRC, resistance, compliance, tidal volume, respiratory rate, Ree, and Xee. A weak association between height and any lung function outcome would suggest that a height adjustment is not required. It is not clinically sensible to make a height adjustment to all outcomes. This idea is supported by the model results shown in Table 4.2 which reinforces the existing clinical understanding that there is no strong association between LCI or t_{PTEF}/t_E values and height development. A height adjustment is therefore not needed for LCI or t_{PTEF}/t_E. All the remaining lung function outcomes exhibited a strong association with height. Among these outcomes, it is expected that the impact of height on lung function would align with the direction observed for age's influence on lung function.

The use of cubic splines terms for age seems justified for each lung function outcome in Table 4.2. The growth and development of lung function are inherently non-linear, with distinct phases of acceleration and deceleration over time, as indicated by Figure 4.5. While the smoothing curves initially suggested a linear relationship for FRC and tidal volume outcomes in Figure 4.5, Figure 4.10 illustrates that the cubic spline models provide a more flexible and nuanced representation of the data, capturing subtle deviations from linearity.

Table 4.2: Effect of log(height) on lung function outcomes: μ Coefficients of log(Height) from Equation 4.16

Variable	Estimate	$e^{Estimate}$	Std. Error	t value	Pr(> |t|)
Functional residual capacity	1.45	4.26	0.052	27.99	<0.001
Lung clearance index	-0.00042	0.99	0.023	-0.018	0.985
Resistance	-1.31	0.27	0.065	-20.066	<0.001
Compliance	1.59	4.90	0.10	15.34	<0.001
t_{PTEF}/t_E	0.12	1.13	0.077	1.58	0.11
Tidal volume	1.34	3.82	0.041	32.43	<0.001
Respiratory rate	-0.48	0.62	0.051	-9.59	<0.001
Ree	-1.31	0.27	0.075	-17.32	<0.001
Xee	-0.22	0.80	0.028	-7.52	<0.001

t_{PTEF}/t_E = Ratio time of peak total expiratory flow to time of expiration; Ree = respiratory resistance at the end of expiration; Xee = respiratory reactance at the end of expiration.
μ link function: log

Visualising the impact of height adjustments on lung function trajectories

The effects of the height adjustment can be visualised in several ways: 1) by creating scatter plots of each lung function outcome against the height adjustment (Y′) itself, 2) by generating scatter plots of each lung function outcome versus height at each annual visit, both before (Y) and after the height adjustment (Y′), and 3) by plotting and comparing the longitudinal trajectories of height-adjusted and unadjusted outcomes.

Figure 4.7 displays the first manner in which the height adjustment can be shown visually by plotting each lung function outcome against the height-adjusted outcome (Y′). Since this approach is similar to the approach used in Figures 4.9a-h, only the FRC and LCI outcomes are used to illustrate the height adjustment here in the main text. The figures which plot resistance, compliance, tidal volume, respiratory rate, Ree, or Xee against their height-adjusted values (Y′) are shown in the Appendix, Figure A.3. Figure 4.7 shows the (a) FRC and (b) LCI measurements plotted against Y′ = FRC/$Height^b$ and Y′ = LCI/$Height^b$. The deviations from a linear trend reflect the adjustments made at each age. These adjustments are particularly noticeable in the FRC outcome, especially during early life, while the LCI demonstrates a linear pattern as anticipated due to the outcome's low association with height.

Figure 4.8 displays the longitudinal trajectories of the height-adjusted outcomes, highlighting cases where this adjustment was required. Therefore, LCI or t_{PTEF}/t_E are not shown here in Figure 4.8. The height adjustment altered the scale of each outcome, however, the overall shape of the trajectory remains consistent with those initially introduced in Figure 4.5.

Figures 4.9a-h plot each lung function outcome against a child's height at each annual visit before (Y) and then again after the height adjustment (Y′). Figure 4.9a highlights FRC vs height, comparing values before and after adjustment. By doing so, the height trend is shown to be removed, as indicated by a more consistent scatter in the Y′ versus Height plots. Figure 4.9b already exhibits a consistent dispersion of LCI points before any height adjustment. Additionally, it is evident that adjusting for height has no discernible impact and is unnecessary in this scenario. As reported before, LCI does not require a height adjustment, Figure 4.9b was included to illustrate this point.

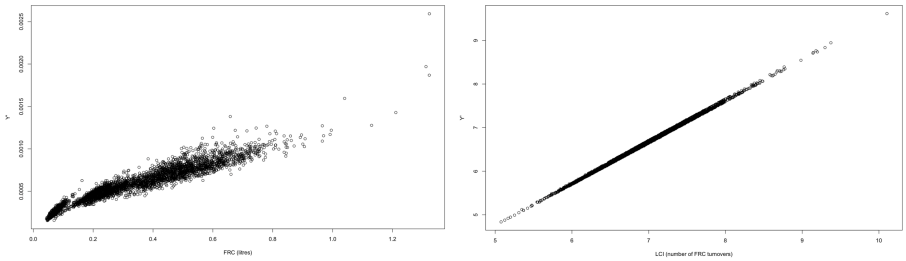


Figure 4.7: FRC measurements against Y′ = FRC/$Height^b$ (left) and LCI measurements against Y′ = LCI/$Height^b$ (right)

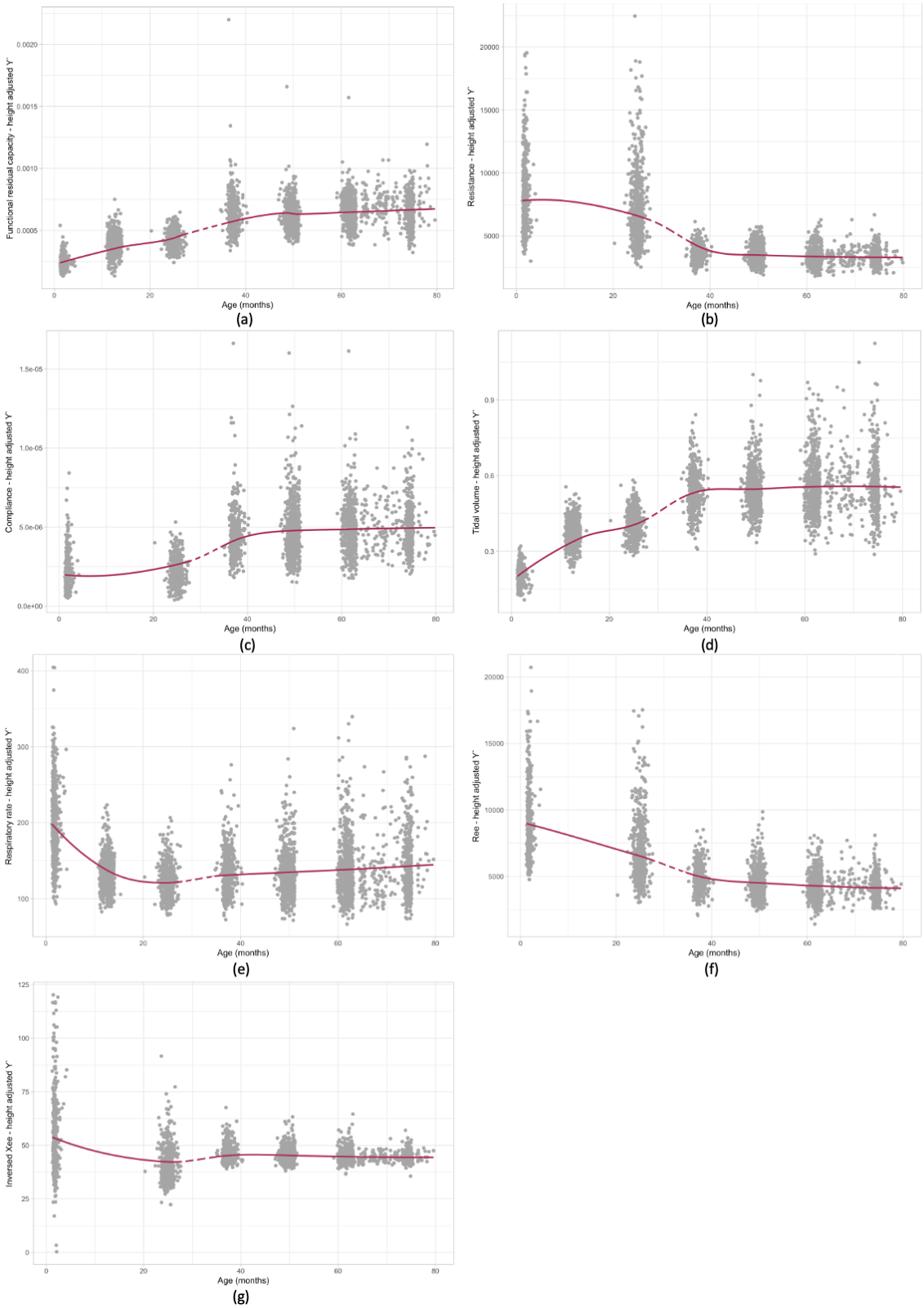
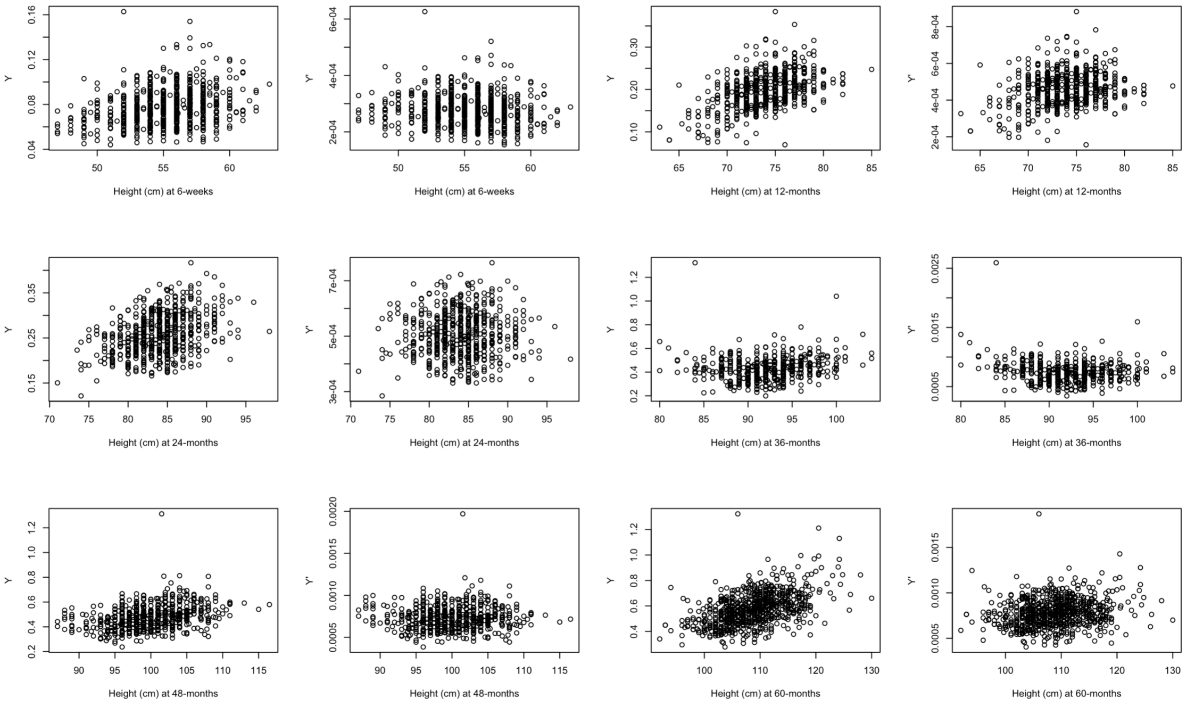
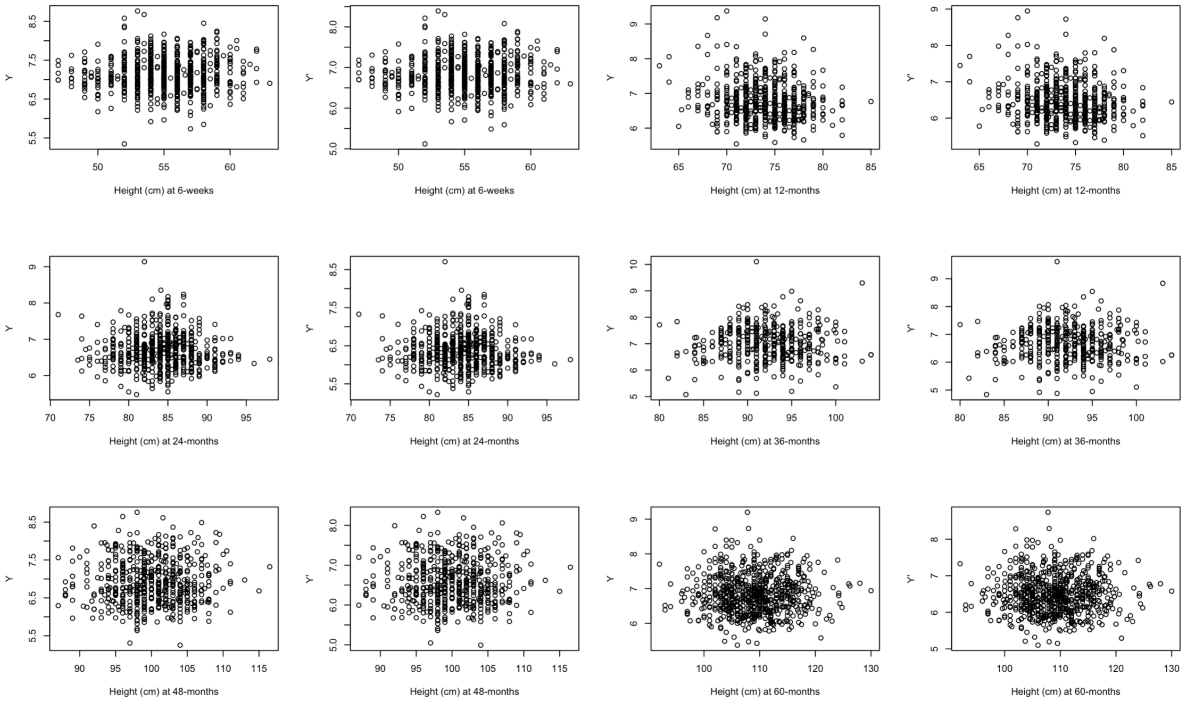


Figure 4.8: Longitudinal plots of height-adjusted lung function outcomes (a) functional residual capacity, (b) resistance, (c) compliance, (d) tidal volume, (e) respiratory rate, (f) respiratory resistance at the end of expiration (Ree), and (g) respiratory reactance at the end of expiration (Xee) versus age. The dotted line between 2 years and 3 years highlight the step change between testing methods.

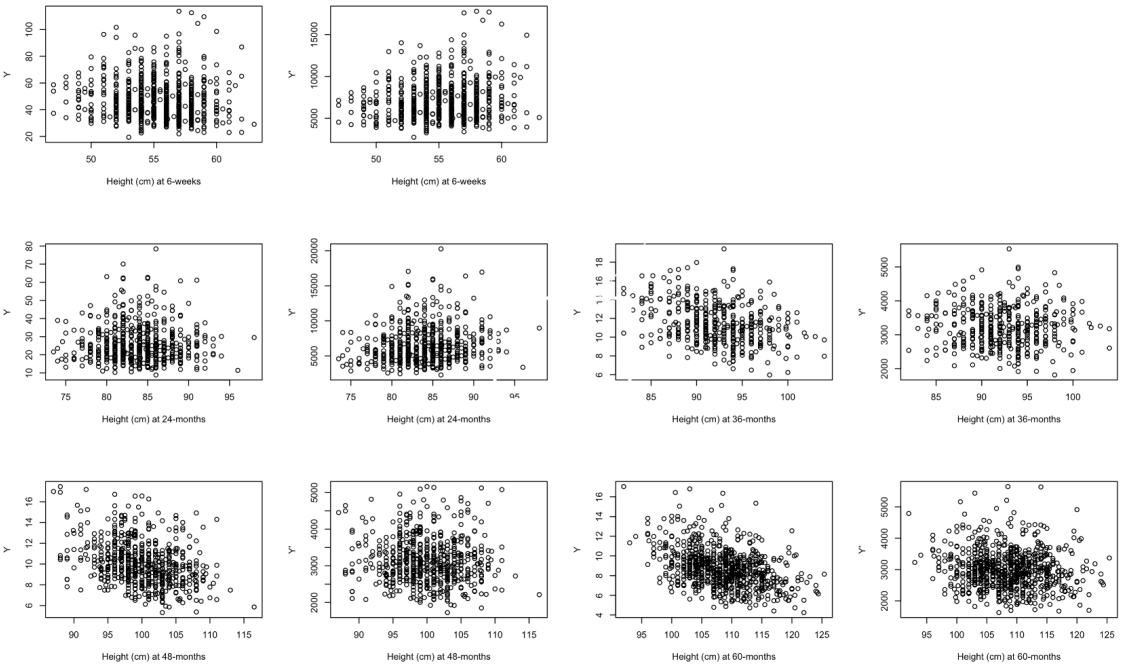


(a)

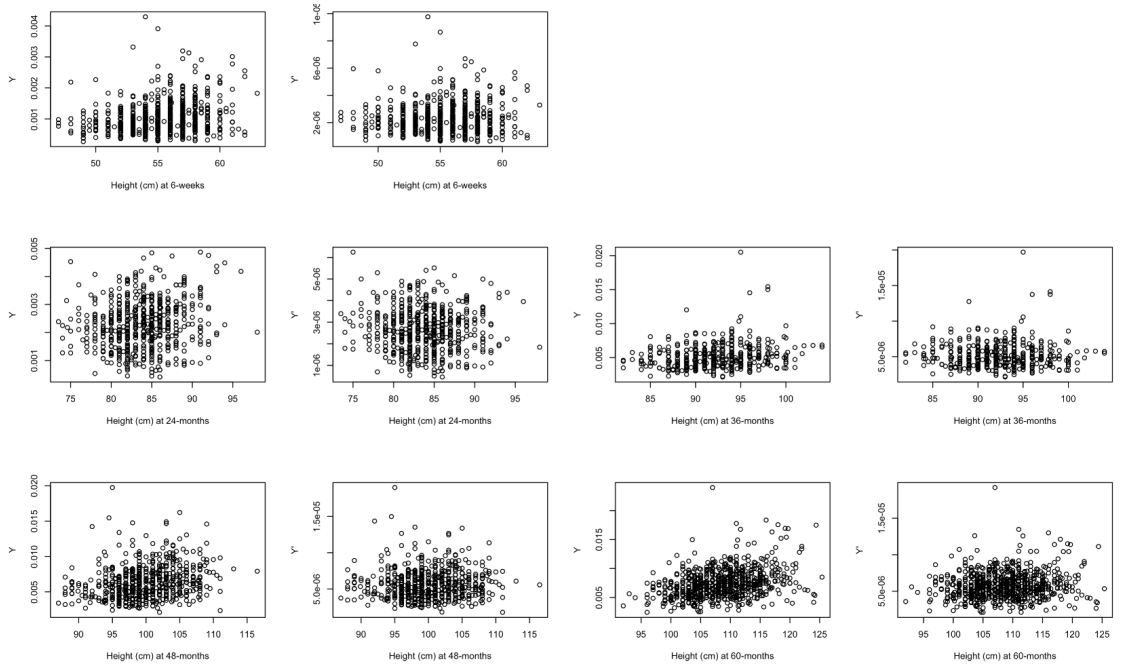


(b)

Figure 4.9: (a) FRC measurements against $Y' = FRC/Height^b$ and (b) LCI measurements against $Y' = LCI/Height^b$. The first and third columns represent unadjusted Y, while the second and fourth columns present height-adjusted Y′.

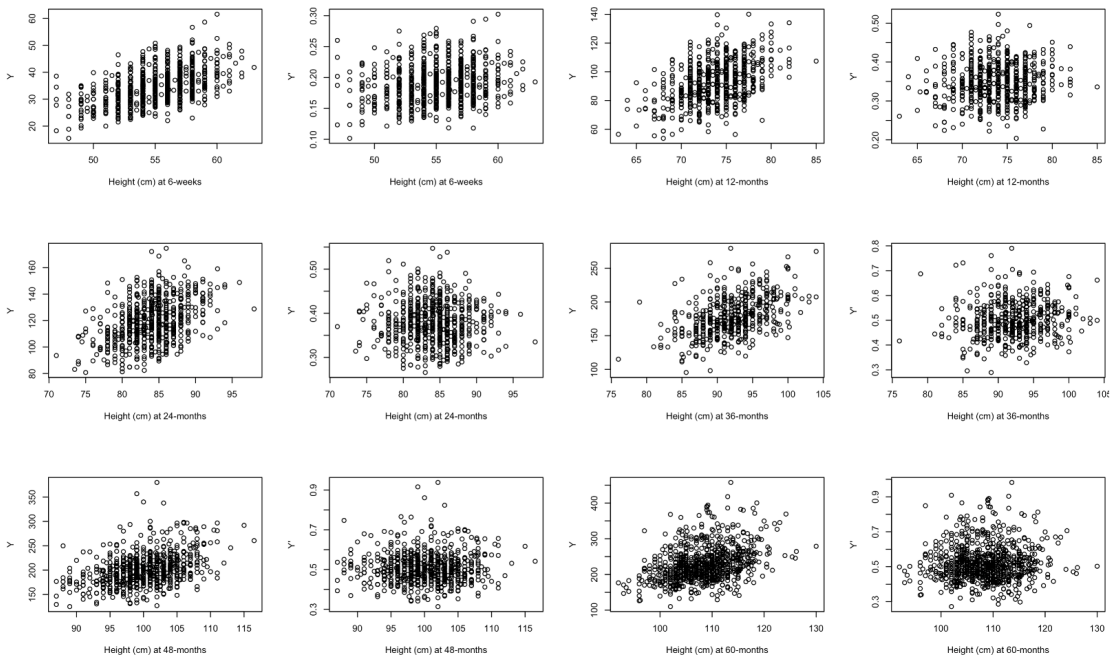


(c)

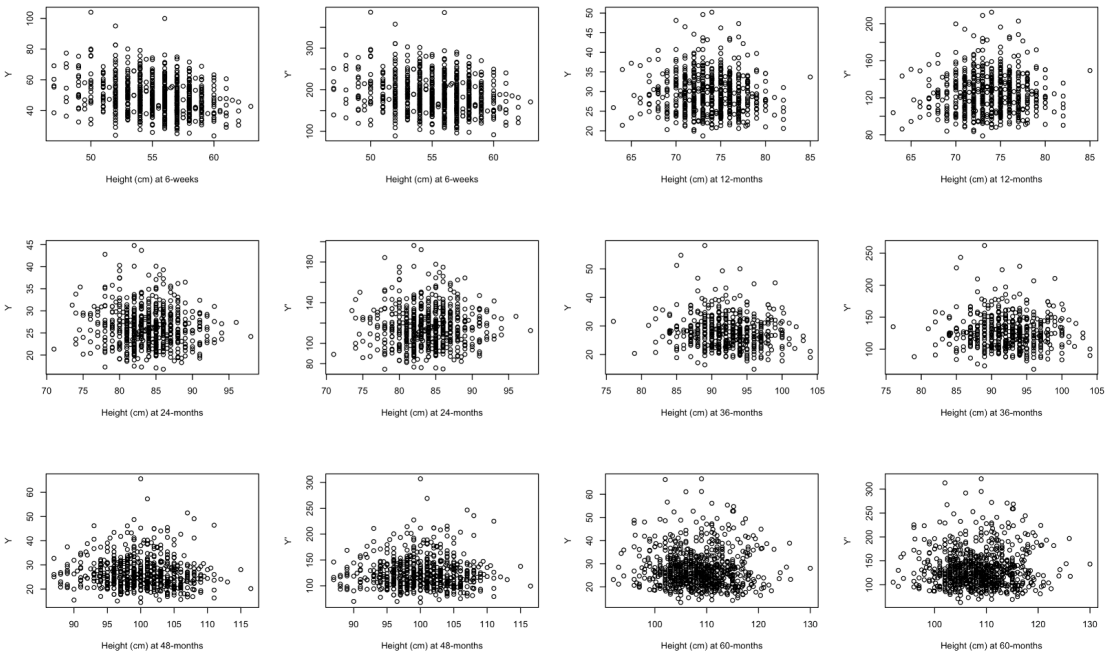


(d)

Figure 4.9: (c) Resistance measurements against $\mathrm{Y}' = \text{Resistance}/Height^b$ and (d) Compliance measurements against $\mathrm{Y}' = \text{Compliance}/Height^b$. The first and third columns represent unadjusted Y, while the second and fourth columns present height-adjusted Y′.

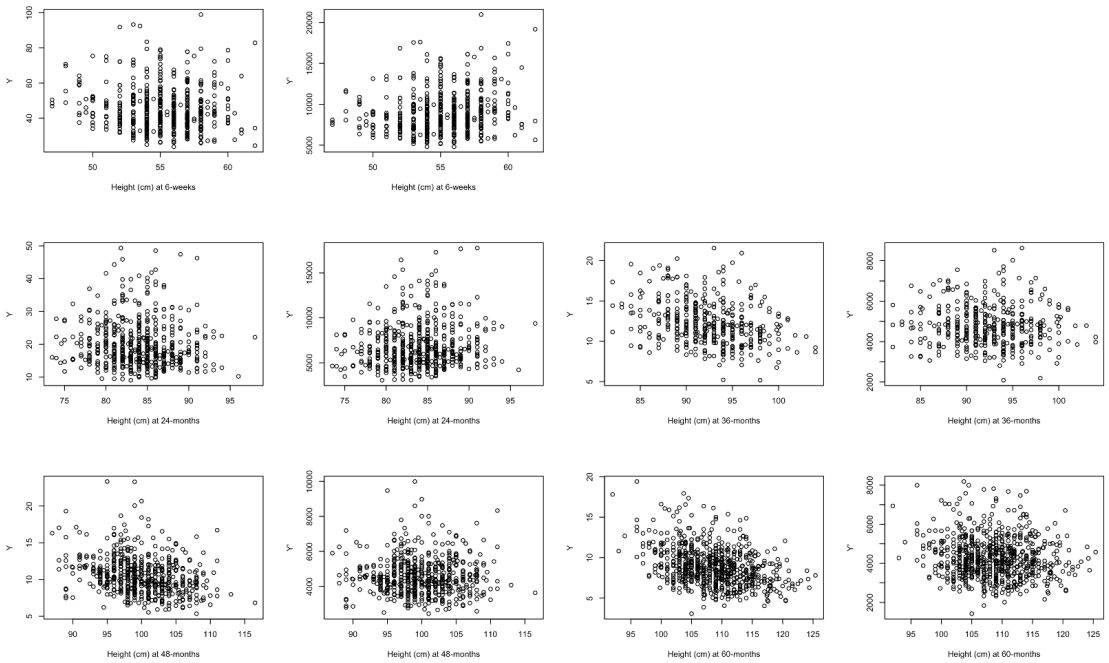


(e)

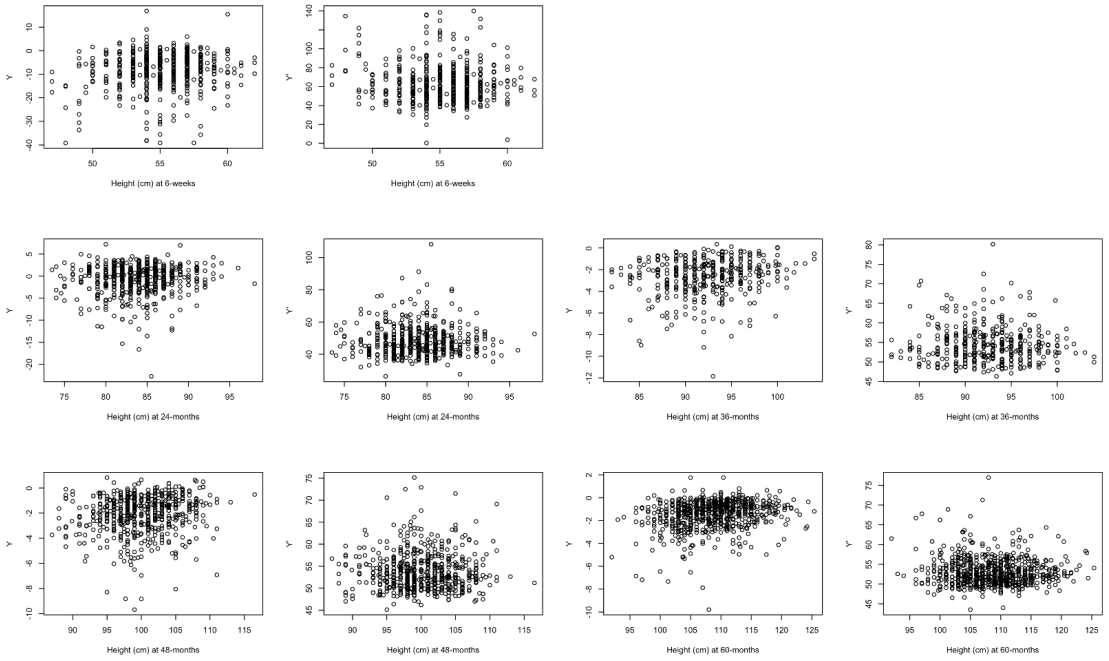


(f)

Figure 4.9: (e) Tidal volume measurements against Y′ = Tidal volume/$Height^b$ and (f) Respiratory rate measurements against Y′ = Respiratory rate/$Height^b$. The first and third columns represent unadjusted Y, while the second and fourth columns present height-adjusted Y′.



(g)



(h)

Figure 4.9: (g) Ree measurements against Y′ = Ree/$Height^b$ and (h) Xee measurements against Y′ = Xee/$Height^b$. The first and third columns represent unadjusted Y, while the second and fourth columns present height-adjusted Y′.

Assessing model estimate consistency with height-adjusted and unadjusted approaches

It is informative to illustrate the implications of incorporating a height adjustment on model coefficients. This section investigates the consistency of the model estimates obtained from modelling either the height-adjusted outcome or the original unadjusted outcome.

Ideally, while modelling the height-adjusted outcome, excluding the height variable from the right-hand side of the regression equation should produce nearly identical coefficients as modelling the original-scale outcome with the height variable included in the regression equation. This comparison allows for an assessment of the consistency in model estimates between height-adjusted and unadjusted approaches.

By using the two models which are introduced in Table 4.3, it can be demonstrated how modelling a height-adjusted outcome produces approximately the same model coefficients as modelling the outcome on the original scale while including the child's height on the right side of the regression equation as an independent variable.

Using the FRC outcome, this example illustrates how the model coefficients are approximately identical between model 1 and model 2 in Table 4.3. In this scenario, both models were fit using the 3-parameter BCCG distribution. Model 2 differs from model 1 by excluding the $log(height)$ term and modelling the height-adjusted FRC measurements. The μ coefficients of model 1 and model 2 are given in Table 4.4 and illustrate the similarity between the model coefficients.

This consistency demonstrates that the adjustment for height in the outcome does not notably alter the model coefficients, validating the equivalence of the two approaches.

Table 4.3: Model formulation used to illustrate the similarity of coefficients while modelling the unadjusted or the height-adjusted FRC outcomes

Model	Distribution	Parameter	Model formulation
1	BCCG	μ	FRC = cs(log(age)) + log(Height) + cs(log(age))*X_t
		σ	σ = cs(log(age))
		ν	ν = cs(log(age))
2	BCCG	μ	Height-adjusted FRC = cs(log(age)) + cs(log(age))*X_t
		σ	σ = cs(log(age))
		ν	ν = cs(log(age))

All models include a child-specific random effect on μ.

In $cs(log(age))$, the $cs()$ term represents a non-parametric cubic spline functions of the log of age.

Age is shown in months.

X_t: a dummy variable indicating the different time periods (< 3 years or > 3 years).

BCCG: Box-Cox Cole and Green distribution.

Table 4.4: Comparison of model coefficients between unadjusted FRC (Model 1) and height-adjusted FRC outcomes (Model 2) from Table 4.3

	FRC				Height-adjusted FRC			
	Estimate	Std. Error	t value	P-value	Estimate	Std. Error	t value	P-value
Intercept	-8.3444	0.2152	-38.760	<0.001	-8.3444	0.0079	-1047.900	<0.001
cs(log(age))	0.2408	0.0091	26.353	<0.001	0.2408	0.0032	73.302	<0.001
log(Height)	1.4072	0.0548	25.647	<0.001	/	/	/	/
X_t	0.5528	0.0693	7.980	<0.001	0.5528	0.0639	8.643	<0.001
cs(log(age)):X_t	-0.0931	0.0182	-5.104	<0.001	-0.0931	0.0164	-5.675	<0.001

4.3.4 Modelling the height-adjusted lung function outcomes

Before exploring different latent trajectories of lung function development, it is helpful to consider the case where there is a singular trajectory. To achieve this, a mean trajectory of lung function development can be modelled to provide insight into what the average development of each outcome looks like. In this section, the height-adjusted values (Y') for each lung function outcome will be analysed to investigate the development of each outcome. When height adjustment is unnecessary, such with LCI or t_{PTEF}/t_E, the outcome will be modelled on the original scale instead of using the height-adjusted data.

The height-adjusted values (Y') for each lung function outcome were modelled as functions of age with subject-specific random effects using a GAMLSS framework. This approach allows flexibility in modelling the response distribution while accounting for within-subject variability due to repeated measurements. The GAMLSS framework is particularly suited for modelling longitudinal data where the mean and other aspects of the distribution, such as variance or skewness, may change over time. By including both fixed and random effects, the model captures both population-level trends and individual-specific deviations.

Understanding and handling outliers

Comprehensive quality control of the data was conducted on the data. This included checking source data for quality, reviewing relevant clinical information at the time of testing and an individual's comprehensive lung function trend over time. All data points included in this analysis are within a well-established clinical range, ensuring its validity. Data that was biologically implausible or on review and did not meet the strict quality definitions were excluded. All other data points were kept and aligned with clinical reasonability. It is acknowledged that certain points could be considered to be outliers as they are associated with a large residual, however, these points are not considered influential observations as they do not have a large impact on model coefficients. The emphasis lies in modelling the mean developmental trajectories in an observational population cohort (children with potentially "unhealthy trajectories" are not excluded), rather than calculating reference ranges or centile curves, where outliers can exert a disproportionate influence, potentially pulling the fitted curves towards the outlier value. Given the overall sample size, these points are not considered to have a large impact on model coefficients. It is however acknowledged that including outlying points would negatively impact the tails of residual plots, such as the normal quantile-quantile plot. A point to note in this analysis is that after adjusting for height, outliers may become magnified. This is evident in the FRC values shown in Figure 4.10.

Modelling trajectories of lung function

The GAMLSS framework was used in combination with an ITS approach to investigate the early-life development of lung function. This enabled the modelling of changes in the age-related methodologies on lung function development over time, accounting for the impact of a 'step-change' which occurred when testing changed from asleep to awake measurements. Models were developed using a 3 parameter (μ, σ, and ν) Box-Cox Cole and Green (BCCG) distribution and a 4 parameter (μ, σ, ν and τ) Box-Cox t (BCT) distribution with increasing complexity, Table 4.8. The decision to implement the structure in Table 4.8 was made based on parsimony, while also emphasising the variation on the data points along the y-axis during each yearly examination, displayed by Figures 4.5 and 4.8. Differences in variation were observed before and after three years. The best-fitting model for each outcome was selected using the GAIC. The height-adjusted outcomes are modelled when necessary, as per Table 4.2. Although a 3-parameter or a 4-parameter distribution was utilised, the primary clinical focus centres on the μ parameter. With reference to the μ parameter, the models presented in Table 4.8 focus on modelling lung function as a function of age and the step variable X_t, without incorporating other risk factors from the dataset just yet. This approach first establishes a baseline model to capture the natural progression of lung function, ensuring a clear understanding of the underlying trend. Additional risk factors are introduced from Section 4.3.5 onwards.

Theoretical densities of the Normal, BCCG, and BCT distributions were plotted over the observed density histograms at each visit to establish which distributions would provide a good fit to the data, shown in the Appendix, Figure A.2. The BCCG and BCT distributions appeared to fit the data the best, especially towards the tails of the distributions, which suggests the need for modelling skewness and kurtosis. The BCT distribution is expected to perform better when dealing with outlying data points due to its thicker tails compared to the standard normal distribution. This is expected to result in improved normal quantile-quantile plots, reflecting a more accurate representation of the data distribution.

Longitudinal lung function outcomes were modelled from birth to school age. The investigation focused on understanding the impact of age and its interaction with a dummy variable, X_t, on each lung function outcome. The Xee outcome was inverted and a constant was added in order to model the outcome using a BCCG and BCT distribution under the GAMLSS framework. Prior to inverting and adding a constant, the Xee data contained negative values which were represented by a left skewed distribution. Lung function outcomes were modelled using a log link for the μ parameter. The results are reported as exponentiated coefficients for μ on the log scale with 95% confidence intervals (CIs).

Using one outcome as an example, the following model was constructed to analyse the development of FRC. The mathematical representation of the model is as follows: Let Y' represent the height-adjusted FRC measurements, where Y' follows a BCT distribution using a log link function for parameter μ of the model. Using the four parameters (μ, σ, ν and τ) from the BCT distribution, the model can be written in the form:

$$\mathrm{Y}' = \begin{cases} log(\mu) = \beta_0 + \beta_1 Age + \beta_2 X_t + \beta_3 Age \cdot X_t + (1|subject) \\ log(\sigma) = \beta_0 + \beta_1 Age + \beta_2 X_t + \beta_3 Age \cdot X_t \\ \nu = \beta_0 + \beta_1 Age \\ log(\tau) = \beta_0 + \beta_1 Age, \end{cases} \tag{4.23}$$

where $1|subject$ represents the random effect for the subject. When modelled using a logarithmic link function, the exponentiated coefficients from model 4.23 (or similar models) show how changes in the predictor variables relate to changes in the outcome variable on a multiplicative scale. For a specific coefficient β_1, exponentiating it (e^{β_1}) gives you the multiplicative factor by which the outcome variable changes for a one-unit increase in the corresponding predictor variable (in this case, Age).

Table 4.5: GAMLSS model summary for height-adjusted FRC (linear model)

		Estimate	$e^{Estimate}$	Std. Error	t value	Pr(> |t|)
μ link function	log					
μ Coefficients	Intercept	-8.32	0.00024	0.0095	-876.38	< 0.001
	Age (months)	0.024	1.024	0.00052	45.81	< 0.001
	X_t	0.843	2.32	0.018	47.57	< 0.001
	Age (months):X_t	-0.022	0.98	0.00059	-37.12	< 0.001
σ link function	log					
σ Coefficients	Intercept	-1.46	0.23	0.038	-38.69	< 0.001
	Age (months)	-0.019	0.98	0.0025	-7.93	< 0.001
	X_t	-0.36	0.69	0.10	-3.51	< 0.001
	Age (months):X_t	0.014	1.014	0.0031	4.62	< 0.001
ν link function	identity					
ν Coefficients	Intercept	-0.31	0.73	0.13	-2.30	0.021
	Age (months)	0.0076	1.0076	0.0044	1.74	0.082
τ link function	log					
τ Coefficients	Intercept	3.38	29.37	0.39	8.63	< 0.001
	Age (months)	-0.027	0.97	0.0066	-4.077	< 0.001
Family	Box-Cox t					
Fitting method	RS()					

From Table 4.5, age is naturally an important predictor of μ. A unitary increase in age produces, on average, a 2.4% increase ($e^{0.024} = 1.0243$) in the height-adjusted FRC outcome. Compared to the first three years, after age three, the slope decreased by an average of approximately 2% ($1 - e^{-0.022} = 0.0217$). Hence, after three years, we continue to observe a positive slope; however, the gradient of the slope is now more gradual. A step-change, of approximately 132.3% ($e^{0.843} = 2.323$) increase in the expected value of the response variable, at three years reflected the different infant and preschool lung function testing methodologies. The scale parameter σ changes with age. Skewness (ν) and kurtosis (τ) are both associated with age, although less so for ν.

The effect of age on lung function development is shown in Table 4.6 (a) and (b). The impact of

age demonstrated the step change in most measures with the change in testing technique at age three. Lung volumes and compliance increased while resistance and respiratory rate decreased with age. Changes were most marked through early life, with decreased slope changes thereafter, Table 4.6 (a) and (b). For example, compared to the first three years, after age three, the slope decreased by an average of approximately 2% for FRC (0.978; 95% CI 0.977-0.979) and tidal volume (0.975; 95% CI 0.974-0.977) and by 0.5% for compliance (0.995; 95% CI 0.993-0.997). Similarly, the rate in respiratory rate (1.016; 95% CI 1.015-1.017) and Ree (1.009; 95% CI 1.007-1.010) decrease was slower beyond 3 years.

Table 4.6: Trajectories of multiple breath washout, tidal breathing and oscillometry outcomes from birth

(a) Multiple breath washout and tidal breathing trajectories

	Functional residual capacity[1]		Lung clearance index[1]		Respiratory rate[2]		Tidal volume[1]		t_{PTEF}/t_E[1]	
	μ coef (95% CI)	P-value	μ coef (95% CI)	P-value	μ coef (95% CI)	P-value	μ coef (95% CI)	P-value	μ coef (95% CI)	P-value
Age (months)	1.024 (1.023-1.025)	< 0.001	0.997 (0.997-0.997)	< 0.001	0.986 (0.985-0.986)	< 0.001	1.026 (1.025-1.027)	< 0.001	0.986 (0.985-0.987)	< 0.001
X_t	2.323 (2.245-2.407)	< 0.001	1.002 (0.979-1.027)	0.843	0.744 (0.717-0.771)	< 0.001	2.353 (2.28-2.429)	< 0.001	1.316 (1.255-1.379)	< 0.001
Age (months):X_t	0.978 (0.977-0.979)	< 0.001	1.002 (1.002-1.003)	< 0.001	1.016 (1.015-1.017)	< 0.001	0.975 (0.974-0.977)	< 0.001	1.011 (1.009-1.012)	< 0.001

(b) Oscillometry trajectories

	Resistance[2]		Compliance[1]		Ree[1]		Inversed Xee[2]	
	μ coef (95% CI)	P-value	μ coef (95% CI)	P-value	μ coef (95% CI)	P-value	μ coef (95% CI)	P-value
Age (months)	0.996 (0.994-0.997)	< 0.001	1.008 (1.006-1.01)	< 0.001	0.987 (0.986-0.988)	< 0.001	0.988 (0.987-0.989)	< 0.001
X_t	0.533 (0.514-0.552)	< 0.001	2.117 (2.002-2.237)	< 0.001	0.627 (0.603-0.652)	< 0.001	0.873 (0.847-0.899)	< 0.001
Age (months):X_t	1.002 (1-1.003)	0.0342	0.995 (0.993-0.997)	< 0.001	1.009 (1.007-1.01)	< 0.001	1.011 (1.01-1.013)	< 0.001

Table notes:
All models include a child-specific random effect on μ.
All lung function measurements except LCI and t_{PTEF}/t_E were standardised for height at test.
X_t: a dummy variable indicating the different time periods (< 3 years or > 3 years)
Ree = respiratory resistance at the end of expiration; Xee = respiratory reactance at the end of expiration.
[1]BCT: Box-Cox t distribution
[2]BCCG: Box-Cox Cole and Green distribution

With regards to model complexity, linear lines may in some cases be an oversimplification of reality, or at least oversimplify particular time stages of development. The section which follows investigates the possible need for increased model complexity.

Increasing model complexity

Models of increasing complexity were evaluated to determine the most appropriate fit for the data. In this analysis, increased complexity is achieved by introducing a quadratic term on the μ component and then further by using a cubic B-spline basis expansion. These more complex models were fit in addition to the linear models. It is important to be mindful of overfitting when using more complicated (using more degrees of freedom) model formulations, this is especially true within the GAMLSS framework. Overfitting can be prevented indirectly by using the GAIC, which imposes a penalty for models with a larger number of parameters, [160]. Table 4.7 displays the model fit statistics for models of increasing complexity for each outcome. The spline models provided the best fit for most lung function outcomes, as indicated by their lower GAIC values. However, for resistance and Ree, the linear models outperformed, offering the optimal fit. While the quadratic model provided the best fit for the Xee outcome.

The fitted trends of the linear, quadratic, and spline models are displayed in Figures 4.10 and 4.11. For the MBW and tidal breathing measures, Figure 4.10 (a)-(e), two key visual differences stand out between the linear, quadratic, and spline models. Firstly, the spline models provide a more accurate fit, particularly over the 12-month measurement mark for tidal volume and respiratory rate, as illustrated by the purple line. Secondly, the quadratic and spline models produce a more flexible fit towards the tails at lower and upper age ranges, while it is worth noting the sparseness of the data points around the upper age range. While spline models offer a smoother fit, they come with the trade-off of a more complex interpretation of model output. In Figure 4.11 (a)-(d), the fitted linear, quadratic, and spline models for the airway oscillometry measurements show minimal visual differences. Given the absence of measurements at 12 months, a linear model seems particularly suitable for the oscillometry measurements.

To summarise, a linear line generally fits the data well after 3 years and before 3 years if we are missing the 12-month measurements (drawing a straight line between 2 points is most reasonable). Non-linearity proved important in two instances with the DCHS cohort: during the initial two years of life and towards the conclusion of the observational period.

To uphold the principle of parsimony and make clinical results easier to interpret, simple linear models may be preferred over more complex non-linear alternatives.

The initial models presented in Table 4.6 offer a simplified per-

	Model	df	GAIC
FRC	Spline	602.15	-61640.76
	Quadratic	537.66	-60932.87
	Linear	533.44	-60892.28
LCI	Spline	260.85	5831.03
	Quadratic	258.12	5874.75
	Linear	249.17	5912.69
RR	Spline	625.47	45219.48
	Quadratic	589.17	45561.05
	Linear	560.92	45728.84
TV	Spline	611.03	-14232.49
	Quadratic	498.04	-12812.91
	Linear	509.89	-12697.43
t_{PTEF}/t_E	Spline	517.59	31505.40
	Quadratic	509.83	31559.88
	Linear	506.90	31567.00
Resist.	Spline	590.44	54880.50
	Quadratic	591.49	54879.28
	Linear	588.10	54877.77
Compl.	Spline	560.71	-81807.41
	Quadratic	912.38	-81032.33
	Linear	555.95	-81798.82
Ree	Spline	480.38	50135.29
	Quadratic	479.30	50134.47
	Linear	478.73	50133.92
Xee	Spline	397.47	17974.29
	Quadratic	396.70	17972.17
	Linear	398.99	17976.12

Table 4.7: Model fit statistics for lung function outcomes

spective, as they do not yet account for the influence of exposures to different risk factors on lung function development. The sections which follow extend these models by incorporating potential risk factors that could shift the developmental trajectories of a child up or down.

Table 4.8: Fitted models for each distribution assumed in GAMLSS

Model	Distribution	Parameter	Model formulation
1	BCCG	μ	$\mu = \beta_0 + \beta_1 Age + \beta_2 X_t + \beta_3 Age X_t$
		σ	$\sigma = \beta_0 + \beta_1 Age$
		ν	$\nu = \beta_0 + \beta_1 Age$
2	BCCG	μ	$\mu = \beta_0 + \beta_1 Age + \beta_2 X_t + \beta_3 Age X_t$
		σ	$\sigma = \beta_0 + \beta_1 Age + \beta_2 X_t + \beta_3 Age X_t$
		ν	$\nu = \beta_0 + \beta_1 Age$
3	BCCG	μ	$\mu = \beta_0 + \beta_1 Age + \beta_2 X_t + \beta_3 Age X_t$
		σ	$\sigma = \beta_0 + \beta_1 Age + \beta_2 X_t + \beta_3 Age X_t$
		ν	$\nu = \beta_0$
4	BCT	μ	$\mu = \beta_0 + \beta_1 Age + \beta_2 X_t + \beta_3 Age X_t$
		σ	$\sigma = \beta_0 + \beta_1 Age$
		ν	$\nu = \beta_0$
		τ	$\tau = \beta_0$
5	BCT	μ	$\mu = \beta_0 + \beta_1 Age + \beta_2 X_t + \beta_3 Age X_t$
		σ	$\sigma = \beta_0 + \beta_1 Age + \beta_2 X_t + \beta_3 Age X_t$
		ν	$\nu = \beta_0$
		τ	$\tau = \beta_0$
6	BCT	μ	$\mu = \beta_0 + \beta_1 Age + \beta_2 X_t + \beta_3 Age X_t$
		σ	$\sigma = \beta_0 + \beta_1 Age + \beta_2 X_t + \beta_3 Age X_t$
		ν	$\nu = \beta_0 + \beta_1 Age$
		τ	$\tau = \beta_0 + \beta_1 Age$
7	BCT	μ	$\mu = \beta_0 + \beta_1 Age + \beta_2 X_t + \beta_3 Age X_t$
		σ	$\sigma = \beta_0 + \beta_1 Age + \beta_2 X_t + \beta_3 Age X_t$
		ν	$\nu = \beta_0$
		τ	$\tau = \beta_0 + \beta_1 Age$
8	BCT	μ	$\mu = \beta_0 + \beta_1 Age + \beta_2 X_t + \beta_3 Age X_t$
		σ	$\sigma = \beta_0 + \beta_1 Age + \beta_2 X_t + \beta_3 Age X_t$
		ν	$\nu = \beta_0 + \beta_1 Age$
		τ	$\tau = \beta_0$

Table notes:
All models include a child-specific random effect on μ.
X_t: a dummy variable indicating the dierent time periods (< 3 years or > 3 years)
BCT: Box-Cox t distribution
BCCG: Box-Cox Cole and Green distribution

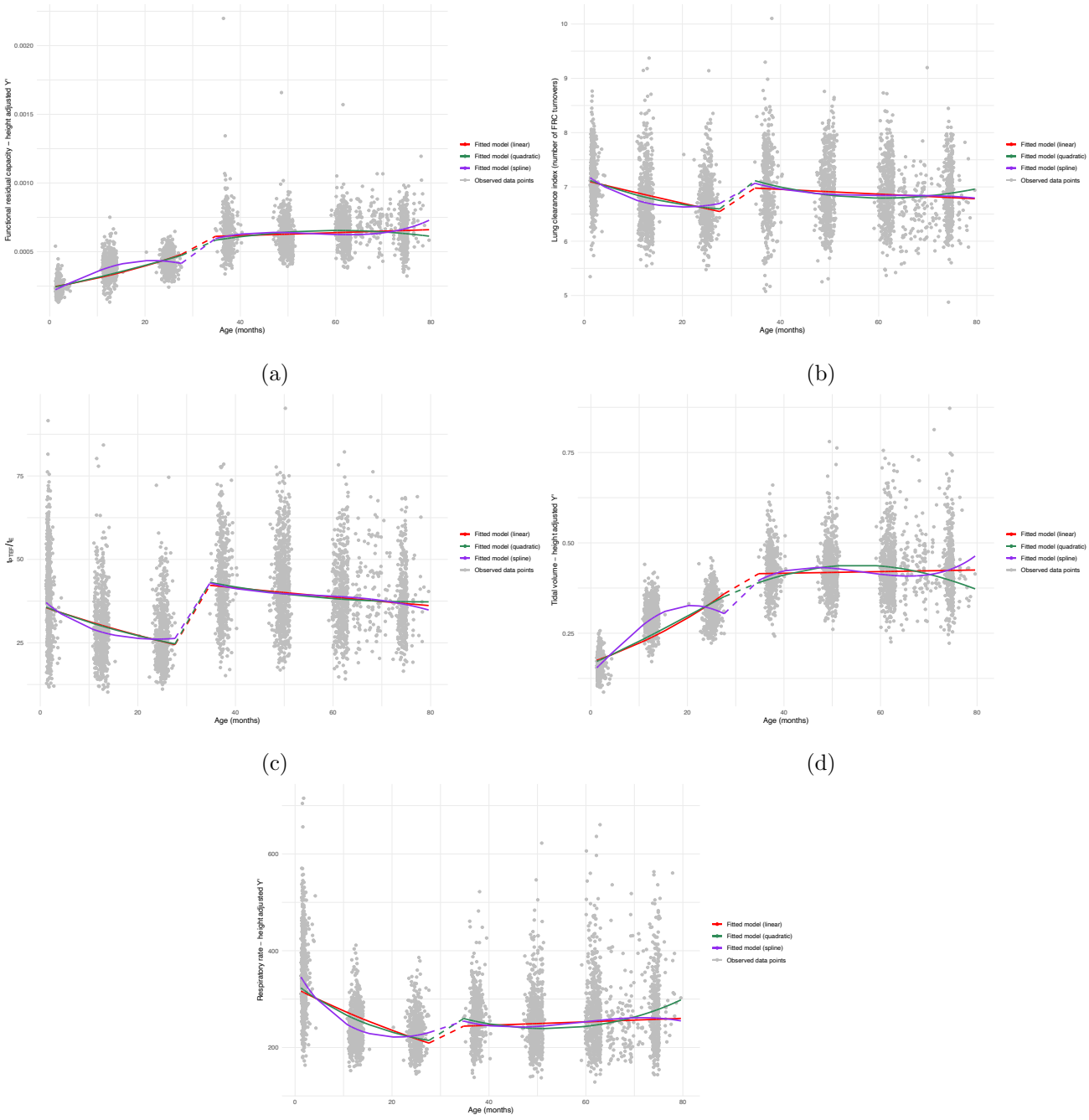
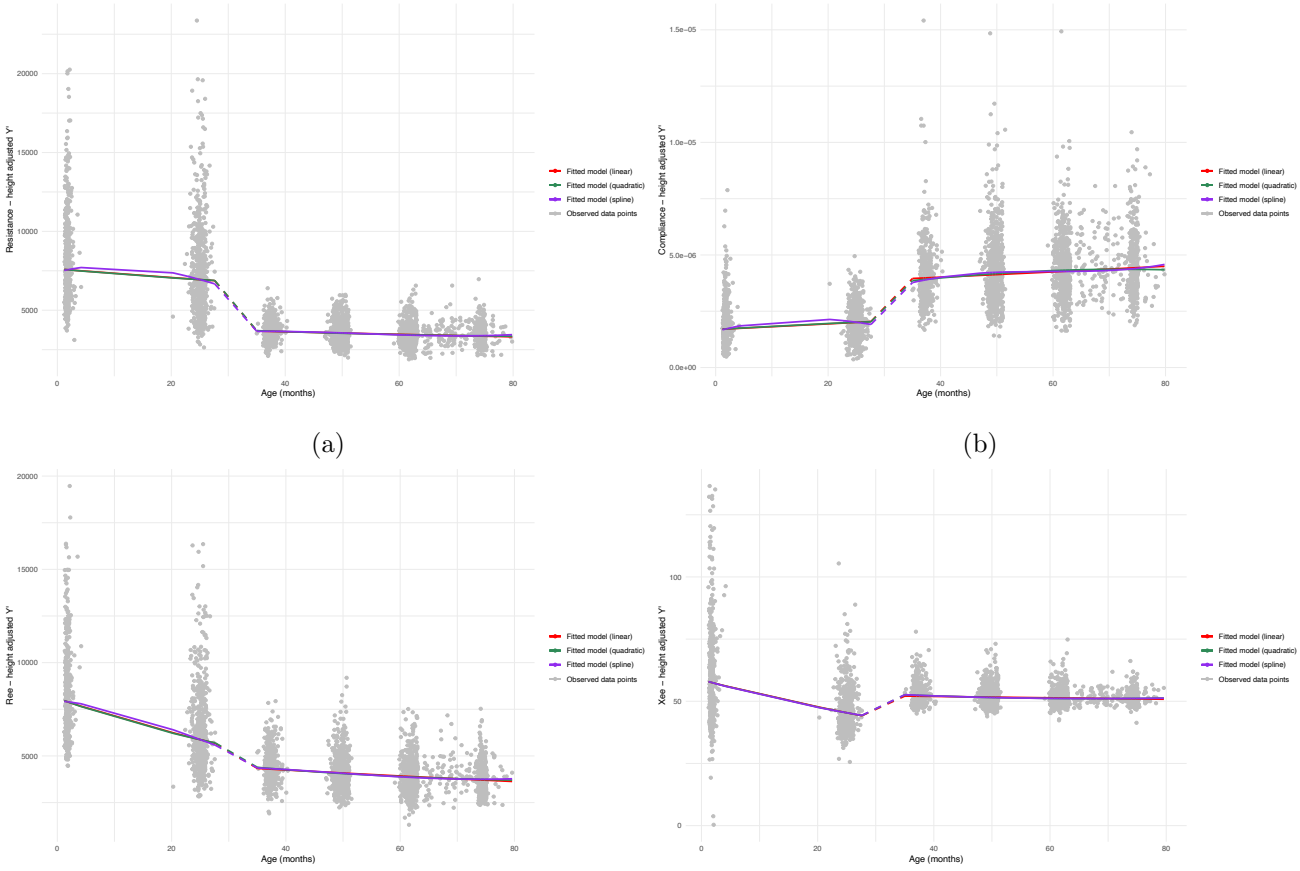


Figure 4.10: Observed data points alongside fitted linear, quadratic, and spline models showcasing each model's ability to represent the underlying pattern. For each measured lung function outcome (a) functional residual capacity, (b) lung clearance index, (c) Ratio time of peak total expiratory flow to time of expiration (t_{PTEF}/t_E), (d) tidal volume, (e) respiratory rate versus age.



(a) (b)

(c) (d)

Figure 4.11: Observed data points alongside fitted linear, quadratic, and spline models showcasing each model's ability to represent the underlying pattern. For each measured lung function outcome (a) resistance, (b) compliance, (c) Ree, and (d) Xee versus age.

4.3.5 Determinants of lung function development

Background

This section investigates the impact of early life LRTI and associated risk factors shaping the development of lung function outcomes from birth to school age, as reported and published by McCready et al. (2024), [157]. The goal is to identify both factors that are detrimental to development and those that are protective. Linear models were prioritised over intricate non-linear models in order to uphold the principle of parsimony and to enhance the ease of interpreting clinical results.

As before, models were developed using a 3 parameter (μ, σ, and ν) BCCG distribution and a 4 parameter (μ, σ, ν, and τ) BCT distribution with increasing complexity. A minimum set of confounders was identified using a DAG as shown in Figure 4.12. Some variables have no direct impact on the lung function outcome but do influence wheeze and LRTI. For each lung function outcome, the investigation focused on a model that included the DAG-identified set of variables: sex, prematurity, wheeze phenotypes, previous LRTI, weight-for-age z-score (at birth), SES components, maternal smoking, maternal HIV, maternal asthma or allergy, and intimate partner violence (IPV). All models additionally contained age, a constant (X_t) representing the size of the step-change at 3 years, an age: X_t interaction term, and a child-specific random effect. As before, lung function outcomes were modelled using a log link for the μ parameter. Results are reported as exponentiated coefficients for μ on the log scale with 95% CIs.

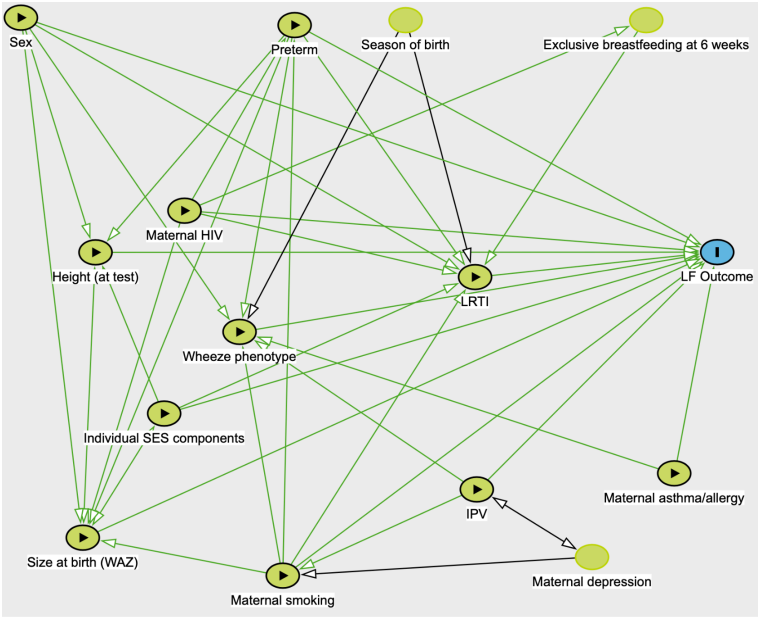


Figure 4.12: DAG for evaluation of exposures.

The best fitting model for each outcome was selected using the GAIC. To investigate the associations

between risk factors and various lung function outcomes, the μ component from the equations introduced in Table 4.8 is extended to incorporate these risk factors. For example, the μ component from model 1 in Table 4.8 is extended as shown in Equation 4.24.

$$\text{Y'} = \begin{cases} log(\mu) = \beta_0 + \beta_1 Age + \beta_2 X_t + \beta_3 Age \cdot X_t + predictor + (1|subject) \\ log(\sigma) = \beta_0 + \beta_1 Age + \beta_2 X_t + \beta_3 Age \cdot X_t \\ \nu = \beta_0 + \beta_1 Age \end{cases} \tag{4.24}$$

4.3.6 Clinical results

Model results

In this section, the effect of age and exposure to various risk factors on the development of lung function parameters, including p-values adjusted for multiple testing calculated using the Benjamini and Hochberg method[161], is shown in Table 4.9 for multiple breath washout and tidal breathing outcomes and in Table 4.10 for oscillometry outcomes.

Male children had larger lung volumes than female children up to age 5 years, reflecting the larger lung size and early airway dysynapsis (disproportionately larger airways relative to lung size) commonly seen in males.

Children born preterm had lower size-adjusted lung volumes (FRC and tidal volume), compliance, and t_{PTEF}/t_E, up to age 5 years compared with term children. These results are believed to reflect both early impaired lung development and an increased vulnerability to lung damage from additional early-life harm. Higher birthweight-z-score was associated with better lung function, larger height-adjusted volumes, higher expiratory flow ratio and compliance, and lower respiratory rate and resistance - all consistent with sustained good lung and airway growth throughout childhood. Maternal asthma and allergy were associated with higher respiratory rate and lower tidal volume over 5 years.

Compared with children in the never wheeze cluster, early transient wheeze was associated with increased LCI. Late-onset wheeze was associated with higher resistance and a lower respiratory rate. Recurrent wheeze was associated with higher Ree and LCI and lower compliance, t_{PTEF}/t_E, and respiratory rate.

Maternal smoking was associated with higher resistance and decreased t_{PTEF}/t_E and respiratory rate, suggestive of airway disease. Postnatal intimate partner violence was associated with lower t_{PTEF}/t_E and Ree and higher FRC.

All socioeconomic status measures were associated with aspects of altered lung function development. Compared with maternal primary education, maternal tertiary education was associated with higher tidal volume and maternal secondary education was associated with higher resistance.

Compared with low asset ownership, high asset ownership was associated with increased tidal volume, lower respiratory rate, and lower t_{PTEF}/t_E. Larger household size was associated with higher compliance, FRC, tidal volume, and t_{PTEF}/t_E. Children from higher-income households had higher FRC than those from lower-income households.

	FRC[1]		LCI[1]		Respiratory rate[1]		Tidal volume[2]		t_{PTEF}/t_E[1]	
	Effect size (95% CI)	**P-value**	**Effect size (95% CI)**	**P-value**	**Effect size (95% CI)**	**P-value**	**Effect size (95% CI)**	**P-value**	**Effect size (95% CI)**	**P-value**
Age (months)	1.024 (1.023-1.025)	<0.001 (<0.001)	0.997 (0.997-0.997)	<0.001 (<0.001)	0.983 (0.983-0.984)	<0.001 (<0.001)	1.025 (1.025-1.026)	<0.001 (<0.001)	0.985 (0.984-0.987)	<0.001 (<0.001)
X_t	2.133 (2.063-2.205)	<0.001 (<0.001)	1.004 (0.984-1.024)	068 (0.91)	0.720 (0.695-0.745)	<0.001 (<0.001)	2.324 (2.261-2.389)	<0.001 (<0.001)	1.312 (1.249-1.377)	<0.001 (<0.001)
Age (months):X_t	0.980 (0.979-0.981)	<0.001 (<0.001)	1.002 (1.002-1.003)	<0.001 (<0.001)	1.017 (1.016-1.018)	<0.001 (<0.001)	0.976 (0.975-0.977)	<0.001 (<0.001)	1.011 (1.009-1.012)	<0.001 (<0.001)
Child characteristics										
Sex (male)	1.062 (1.051-1.073)	< 0.001 (<0.001)	0.998 (0.993-1.003)	0.39 (0.87)	0.999 (0.989-1.010)	0.98 (0.98)	1.032 (1.023-1.041)	<0.001 (<0.001)	0.991 (0.975-1.007)	0.25 (0.37)
Preterm (<37 weeks)	0.982 (0.968-0.997)	0.016 (0.05)	0.999 (0.993-1.006)	0.87 (0.97)	1.015 (1.0005-1.029)	0.042 (0.062)	0.975 (0.963-0.987)	<0.001 (<0.001)	0.965 (0.944-0.986)	0.002 (0.006)
HIV exposed uninfected	1.0006 (0.987-1.014)	0.93 (0.94)	1.008 (1.002-1.014)	0.016 (0.10)	1.005 (0.991-1.018)	0.49 (0.98)	1.022 (1.011-1.034)	<0.001 (<0.001)	0.996 (0.976-1.017)	0.74 (0.80)
Birthweight-z-score	1.006 (1.001-1.011)	0.018 (0.26)	1.0009 (0.998-1.003)	0.42 (0.86)	0.993 (0.988-0.998)	0.0064 (0.016)	1.013 (1.009-1.017)	<0.001 (<0.001)	1.007 (0.999-1.015)	0.052 (0.10)
Wheeze phenotype										
Early transient (vs Never)	0.994 (0.981-1.008)	0.41 (0.51)	1.009 (1.003-1.016)	0.0043 (0.036)	1.005 (0.992-1.019)	0.44 (0.53)	0.992 (0.981-1.003)	0.17 (0.30)	1.017 (0.996-1.038)	0.11 (0.19)
Late onset (vs Never)	1.017 (0.999-1.034)	0.062 (0.13)	0.995 (0.987-1.004)	0.28 (0.81)	0.978 (0.961-0.995)	0.012 (0.027)	0.992 (0.978-1.007)	0.31 (0.48)	1.011 (0.984-1.038)	0.42 (0.55)
Recurrent (vs Never)	1.012 (0.996-1.028)	0.15 (0.25)	1.008 (1.001-1.015)	0.047 (0.26)	0.980 (0.964-0.996)	0.014 (0.029)	1.006 (0.992-1.019)	0.41 (0.54)	0.960 (0.936-0.984)	0.002 (0.006)
LRTI										
LRTI ever	0.987 (0.975-0.998)	0.022 (0.055)	0.999 (0.994-1.005)	0.91 (0.97)	1.018 (1.0063-1.029)	0.0024 (0.0067)	0.993 (0.984-1.003)	0.18 (0.31)	1.002 (0.984-1.020)	0.79 (0.82)
Maternal characteristics										
Postnatal IPV	1.015 (1.005-1.026)	0.0051 (0.021)	0.997 (0.992-1.002)	0.29 (0.81)	1.005 (0.994-1.015)	0.37 (0.49)	1.001 (0.991-1.009)	0.98 (0.98)	0.982 (0.966-0.998)	0.033 (0.075)
Maternal asthma/allergy	1.016 (0.996-1.036)	0.11 (0.21)	0.998 (0.988-1.007)	0.67 (0.91)	1.021 (1.002-1.041)	0.034 (0.059)	0.986 (0.969-1.002)	0.086 (0.18)	1.030 (0.999-1.061)	0.053 (0.10)
Postnatal smoking	0.984 (0.973-0.996)	0.0092 (0.032)	1.0002 (0.995-1.006)	0.93 (0.97)	0.979 (0.968-0.990)	<0.001 (<0.001)	1.008 (0.999-1.018)	0.084 (0.18)	0.976 (0.959-0.994)	0.0084 (0.021)
Socio-economic status										
Maternal education										
Secondary (vs Primary)	0.986 (0.966-1.005)	0.16 (0.25)	0.997 (0.988-1.006)	0.54 (0.90)	1.023 (1.003-1.044)	0.022 (0.042)	1.012 (0.995-1.029)	0.15 (0.29)	1.003 (0.973-1.034)	0.83 (0.83)
Completed secondary (vs Primary)	0.995 (0.974-1.016)	0.64 (0.71)	0.997 (0.987-1.007)	0.58 (0.91)	1.020 (0.998-1.042)	0.068 (0.094)	1.007 (0.989-1.025)	0.43 (0.54)	1.014 (0.982-1.047)	0.3 (0.54)
Any tertiary (vs Primary)	0.979 (0.951-1.008)	0.15 (0.25)	0.999 (0.986-1.013)	0.97 (0.97)	0.999 (0.971-1.029)	0.98 (0.98)	1.052 (1.026-1.077)	<0.001 (<0.001)	0.970 (0.927-1.015)	0.19 (0.29)
Asset ownership										
Low-Medium (vs Low)	1.003 (0.988-1.018)	0.66 (0.72)	1.001 (0.994-1.008)	0.73 (0.91)	0.984 (0.969-0.999)	0.038 (0.059)	1.0006 (0.988-1.013)	0.92 (0.96)	0.994 (0.971-1.017)	0.61 (0.73)
Medium-High (vs Low)	1.0006 (0.985-1.016)	0.94 (0.94)	1.003 (0.995-1.009)	0.48 (0.86)	0.972 (0.957-0.987)	<0.001 (<0.001)	1.005 (0.991-1.017)	0.48 (0.54)	0.955 (0.932-0.978)	<0.001 (<0.001)
High (vs Low)	1.018 (1.001-1.035)	0.033 (0.075)	0.995 (0.987-1.002)	0.18 (0.64)	0.954 (0.939-0.970)	<0.001 (<0.001)	1.022 (1.008-1.036)	0.002 (0.005)	0.963 (0.938-0.988)	0.004 (0.005)
Household size										
Small-Medium (vs small)	0.993 (0.977-1.008)	0.34 (0.45)	1.005 (0.998-1.012)	0.18 (0.64)	0.990 (0.975-1.005)	0.22 (0.31)	1.011 (0.997-1.024)	0.11 (0.20)	1.004 (0.981-1.028)	0.71 (0.77)
Medium-Large (vs small)	1.028 (1.012-1.045)	<0.001 (<0.001)	1.0002 (0.993-1.008)	0.95 (0.97)	0.989 (0.974-1.005)	0.21 (0.31)	1.019 (1.005-1.033)	0.005 (0.013)	1.017 (0.992-1.042)	0.17 (0.27)
Large (vs small)	1.005 (0.989-1.021)	0.55 (0.65)	1.002 (0.994-1.009)	0.61 (0.94)	0.989 (0.974-1.005)	0.18 (0.28)	0.998 (0.985-1.012)	0.79 (0.86)	1.029 (1.004-1.055)	0.021 (0.058)
Income										
R1 000-5 000 ($67-336) vs (< R1 000 ($67))	1.012 (1.001-1.024)	0.032 (0.068)	0.999 (0.994-1.005)	0.86 (0.97)	1.004 (0.993-1.015)	0.43 (0.52)	1.004 (0.994-1.013)	0.44 (0.54)	0.979 (0.962-0.996)	0.015 (0.25)
< R5 000 ($336) vs < R1 000 ($67)	1.021 (1.003-1.039)	0.019 (0.052)	0.998 (0.989-1.005)	0.43 (0.86)	1.002 (0.984-1.019)	0.85 (0.92)	0.993 (0.979-1.007)	0.33 (0.48)	0.995 (0.969-1.022)	0.73 (0.80)

Table notes:
All models include a child-specific random effect on μ, all variables in the table included in model.
All lung function measurements except LCI and t_{PTEF}/t_E were standardised for height at test.
X_t: a dummy variable indicating the different time periods (< 3 years or > 3 years)
FRC = functional residual capacity; LCI = lung clearance index; tPTEF/tE = time to peak tidal expiratory time over total expiratory time; IPV = intimate partner violence, LRTI = lower respiratory tract infection.
Effect size: exponent of the mu-coefficients (percentage change of lung function measure associated with the predictor variable in question compared to the reference category)
[1]BCT: Box-Cox t distribution
[2]BCCG: Box-Cox Cole and Green distribution
[3]Adjusted for multiple comparisons, an approach by Benjamini, Y., and Hochberg, Y, [161].

Table 4.9: Association of early life risk factors with the development of multiple breath washout and tidal breathing parameters from birth to age 5 years; multivariate model.

	Resistance [1]		Compliance [1]		Ree [1]		Inversed Xee[2]	
	Effect size (95% CI)	**P-value**	**Effect size (95% CI)**	**P-value**	**Effect size (95% CI)**	**P-value**	**Effect size (95% CI)**	**P-value**
Age (months)	0.996 (0.995-0.998)	<0.001 (<0.001)	1.008 (1.006-1.009)	<0.001 (<0.001)	0.987 (0.986-0.987)	<0.001 (<0.001)	0.987 (0.986-0.989)	<0.001 (<0.001)
X_t	0.533 (0.515-0.552)	<0.001 (<0.001)	2.126 (2.014-2.245)	<0.001 (<0.001)	0.621 (0.596-0.647)	<0.001 (<0.001)	0.618 (0.593-0.644)	<0.001 (<0.001)
Age (months):X_t	1.001 (0.999-1.003)	0.14 (0.29)	0.995 (0.993-0.997)	<0.001 (<0.001)	1.009 (1.007-1.010)	<0.001 (<0.001)	1.009 (1.007-1.010)	<0.001 (<0.001)
Child characteristics								
Sex (male)	0.978 (0.967-0.989)	<0.001 (<0.001)	1.051 (1.032-1.069)	<0.001 (<0.001)	0.984 (0.970-0.998)	<0.001 (<0.001)	0.984 (0.970-0.998)	0.033 (0.75)
Preterm (<37 weeks)	1.059 (1.042-1.075)	<0.001 (<0.001)	0.916 (0.894-0.940)	<0.001 (<0.001)	1.062 (1.041-1.084)	<0.001 (<0.001)	1.064 (1.042-1.086)	<0.001 (<0.001)
HIV exposed uninfected versus HIV unexposed	1.011 (0.996-1.026)	0.14 (0.29)	1.003 (0.979-1.027)	0.81 (0.81)	0.998 (0.98-1.018)	0.88 (0.092)	1.004 (0.998-1.009)	0.89 (0.94)
Birthweight-z-score	0.993 (0.987-0.998)	0.011 (0.034)	1.021 (1.013-1.030)	<0.001 (<0.001)	0.995 (0.988-1.002)	0.15 (0.25)	0.994 (0.987-1.001)	0.12 (0.20)
Wheezing phenotype								
Early transient (vs Never)	0.994 (0.979-1.008)	0.38 (0.56)	1.013 (0.989-1.037)	0.27 (0.41)	1.008 (0.989-1.026)	0.41 (0.57)	1.009 (0.991-1.029)	0.31 (0.45)
Late onset (vs Never)	1.026 (1.007-1.045)	0.0068 (0.024)	0.972 (0.943-1.002)	0.065 (0.13)	1.045 (1.021-1.070)	<0.001 (<0.001)	1.048 (1.022-1.073)	0.002 (0.0071)
Recurrent (vs Never)	1.009 (0.992-1.027)	0.28 (0.53)	0.948 (0.922-0.975)	<0.001 (<0.001)	1.039 (1.016-1.063)	<0.001 (<0.001)	1.041 (1.017-1.065)	<0.001 (<0.001)
LRTI								
LRTI ever	1.028 (1.016-1.041)	<0.001 (<0.001)	0.959 (0.941-0.978)	<0.001 (<0.001)	1.034 (1.017-1.051)	<0.001 (<0.001)	1.032 (1.015-1.049)	<0.001 (<0.001)
Maternal characteristics								
Postnatal IPV	0.991 (0.979-1.003)	0.13 (0.29)	1.013 (0.994-1.032)	0.18 (0.32)	0.983 (0.968-0.997)	0.019 (0.047)	0.982 (0.967-0.997)	0.02 (0.05)
Maternal asthma/allergy	0.995 (0.974-1.016)	0.65 (0.77)	1.017 (0.983-1.053)	0.32 (0.41)	0.964 (0.938-0.991)	0.0091 (0.025)	0.966 (0.938-0.993)	0.016 (0.05)
Postnatal smoking	1.029 (1.016-1.042)	<0.001 (<0.001)	1.003 (0.984-1.023)	0.74 (0.80)	1.018 (1.002-1.034)	0.024 (0.054)	1.019 (1.003-1.036)	0.019 (0.05)
Socio-economic status (SES)								
Maternal education								
Secondary (vs Primary)	1.010 (0.989-1.032)	0.34 (0.56)	0.982 (0.949-1.016)	0.31 (0.41)	1.001 (0.975-1.028)	0.92 (0.92)	1.001 (0.974-1.029)	0.94 (0.94)
Completed secondary (vs Primary)	1.027 (1.004-1.049)	0.022 (0.061)	0.984 (0.949-1.021)	0.41 (0.46)	1.024 (0.995-1.054)	0.11 (0.21)	1.024 (0.994-1.054)	0.12 (0.20)
Any tertiary (vs Primary)	0.996 (0.965-1.029)	0.85 (0.88)	0.969 (0.921-1.021)	0.24 (0.40)	0.997 (0.957-1.039)	0.91 (0.92)	0.999 (0.957-1.04)	0.91 (0.94)
Asset ownership								
Low-Medium (vs Low)	1.0008 (0.985-1.017)	0.92 (0.92)	1.012 (0.986-1.038)	0.37 (0.44)	1.012 (0.991-1.033)	0.27 (0.40)	1.010 (0.989-1.032)	0.35 (0.49)
Medium-High (vs Low)	0.997 (0.989-1.015)	0.77 (0.85)	1.034 (1.006-1.063)	0.015 (0.037)	1.003 (0.981-1.024)	0.81 (0.92)	1.002 (0.979-1.024)	0.88 (0.94)
High (vs Low)	1.005 (0.987-1.024)	0.56 (0.77)	0.986 (0.958-1.015)	0.33 (0.41)	1.002 (0.979-1.025)	0.89 (0.92)	1.001 (0.978-1.025)	0.91 (0.94)
Household size								
Small-Medium (vs small)	1.007 (0.991-1.024)	0.38 (0.56)	1.018 (0.992-1.046)	0.18 (0.32)	0.985 (0.965-1.006)	0.17 (0.26)	0.983 (0.962-1.005)	0.13 (0.20)
Medium-Large (vs small)	1.008 (0.991-1.026)	0.35 (0.56)	1.004 (0.976-1.032)	0.79 (0.81)	0.992 (0.970-1.014)	0.47 (0.61)	0.989 (0.967-1.013)	0.38 (0.50)
Large (vs small)	0.995 (0.978-1.013)	0.59 (0.77)	1.046 (1.017-1.075)	0.002 (0.005)	0.979 (0.957-1.001)	0.056 (0.12)	0.977 (0.956-1.000)	0.049 (0.10)
Income								
R1 000-5 000 ($67-336) vs (< R1 000 ($67))	1.003 (0.991-1.015)	0.63 (0.77)	1.010 (0.991-1.029)	0.31 (0.41)	1.004 (0.988-1.019)	0.64 (0.80)	1.002 (0.986-1.018)	0.81 (0.94)
< R5 000 ($336) vs < R1 000 ($67)	0.997 (0.979-1.016)	0.78 (0.85)	1.029 (0.999-1.061)	0.057 (0.13)	0.982 (0.958-1.005)	0.13 (0.23)	0.979 (0.955-1.004)	0.093 (0.18)

Table notes:
All models include a child-specific random effect on μ, all variables in the table included in model.
X_t: a dummy variable indicating the different time periods (< 3 years or > 3 years)
Ree = respiratory resistance at the end of expiration; Xee = respiratory reactance at the end of expiration; IPV = intimate partner violence, LRTI = lower respiratory tract infection.
Effect size: exponent of the mu-coefficients (percentage change of lung function measure associated with the predictor variable in question compared to the reference category)
[1]BCT: Box-Cox t distribution
[2]BCCG: Box-Cox Cole and Green distribution
[3]Adjusted for multiple comparisons, an approach by Benjamini, Y., and Hochberg, Y, [161].

Table 4.10: Association of early life risk factors with the development of oscillometry parameters from birth to 5 years; multivariate model

HIV exposed uninfected

This section takes a closer look at the lung function development of children who were HIV-exposed uninfected (HEU).

It was initially shown in Table 4.9 that children who were HEU had mildly increased LCI (1.008 [1.002-1.014]) compared with HIV-unexposed children. It was also shown in a previous DCHS that lung function at 2 years is most impacted in HEU whose mothers had the most advanced disease perinatally, [162]. The altered lung function in children who were HEU was associated with maternal HIV disease severity, with the increased LCI noted only in infants whose mothers had a detectable HIV viral load during pregnancy, Table 4.11. Whether this is a consequence of the direct or indirect effects of severe maternal disease, is unknown and requires further study.

	LCI[1]		Tidal volume[1]		Inversed Xee[2]	
	Effect size (95% CI)	**P-value**	**Effect size (95% CI)**	**P-value**	**Effect size (95% CI)**	**P-value**
HIV exposed uninfected versus HIV unexposed	1.008 (1.002-1.014)	0.016	1.022 (1.011-1.034)	<0.0001	1.004 (0.998-1.009)	0.12
Maternal Cd4 categories during pregnancy						
($\geq$ 500 cells/mm^3) versus maternal HIV negative	1.006 (0.996-1.016)	0.21	1.038 (1.021-1.056)	<0.0001	1.011 (1.003-1.018)	0.0046
(350-500 cells/mm^3) versus maternal HIV negative	1.005 (0.993-1.017)	0.41	1.006 (0.984-1.029)	0.61	1.008 (0.997-1.019)	0.14
(< 350 cells/mm^3) versus maternal HIV negative	1.009 (0.999-1.019)	0.062	1.011 (0.993-1.029)	0.25	1.011 (1.003-1.019)	0.0059
Maternal Viral load (vl) categories during pregnancy						
(Lower than detectable limit) versus maternal HIV negative	1.006 (0.997-1.014)	0.21	1.011 (0.996-1.027)	0.16	1.017 (1.01-1.025)	<0.0001
(VL detectable $\geq$40-1000 copies/ml) versus maternal HIV negative	1.017 (1.004-1.031)	0.013	1.01 (0.986-1.036)	0.41	1.02 (1.009-1.031)	<0.0001
(Virally unsuppressed >1000 copies/ml) versus maternal HIV negative	1.022 (1.006-1.038)	0.008	1.036 (1.006-1.067)	0.020	1.011 (0.997-1.026)	0.12

Table notes:
The first three estimates for cd4 categories and the last three estimates for viral load categories are from separate models that included age (months) and 3 year testing step change adjustment.
Xee = respiratory reactance at the end of expiration; LCI = lung clearance index; HIV = human immunodeficiency virus.
[1]BCT: Box-Cox t distribution
[2]BCCG: Box-Cox Cole and Green distribution
Effect size: exponent of the mu-coefficients

Table 4.11: Association of maternal HIV disease severity with LCA, tidal volume, and Xee from birth to 5 years.

In a multi-site, international case-control study conducted across Bangladesh, Gambia, Kenya, Mali, South Africa, Thailand, and Zambia, utilising the PERCH cohort, it was found that children who were HEU had a higher attributable fraction of RSV post 1 year of age compared to children who were not HIV-exposed, [163]. Within the DCHS, LRTI was more common in HEU compared to unexposed infants. Although both HEU and HIV-unexposed infants had a high incidence of LRTI, the incidence of RSV-LRTI was not notably higher in the HEU group, as shown in Table 4.12.

	HIV exposed uninfected (n=211)	**HIV unexposed (N=755)**	**p-value**
Any LRTI	121 (57.3%)	363 (48.1%)	0.021
Non-RSV LRTI	79 (37.4%)	240 (31.8%)	0.14
RSV-LRTI	42 (19.9%)	123 (16.3%)	0.26

Table 4.12: Comparison of LRTI incidence between HIV exposed uninfected and HIV unexposed groups

LRTI results

This section offers a deeper analysis of the association between LRTI and lung function outcomes. Compared with children who never had an LRTI, children with a history of LRTI displayed reduced FRC and compliance, as well as higher resistance and respiratory rate up to age 5, independent of risk contributed by baseline lung function, Tables 4.9 and 4.10. These findings are consistent with impaired lung growth.

Next, the association of baseline lung function with subsequent LRTI is shown in Table 4.13. This association has been previously highlighted as important in the DCHS, [11]. The lung function outcomes are presented as medians with interquartile ranges (IQR), and the p-values indicate the strength of the differences between the groups. The table highlights that respiratory rate, t_{PTEF}/t_E, and inverted Xee show the strongest differences between the groups, indicating that these outcomes are affected in children who have had an LRTI. A visual representation of the results in Table 4.13 is shown in the appendix through Figure A.4.

	Never had LRTI ever [median (IQR)]	**Had LRTI ever [median (IQR)]**	**P-value**[1]
Functional residual capacity	0.00027 (0.00024-0.00031)	0.00026 (0.00023-0.00031)	0.064
Lung clearance index	7.09 (6.87-7.42)	7.13 (6.88-7.43)	0.51
Tidal volume	0.19 (0.17-0.21)	0.18 (0.16-0.20)	0.33
Respiratory rate	178.11 (154.35-203.93)	189.51 (162.27-217.20)	<0.001
t_{PTEF}/t_E	39.5 (30.3-46.8)	35.6 (28.3-44.8)	0.0023
Compliance	0.0022 (0.0017-0.0028)	0.0020 (0.0015-0.0026)	0.050
Resistance	6502.52 (5469.26-8036.63)	6763.83 (5534.95-8314.71)	0.11
Inverted Xee	55.11 (47.50-65.59)	61.89 (51.71-75.40)	0.0012
Ree	8308.11 (7141.18-9873.54)	8694.81 (7420.76-10459.66)	0.21

Table notes:
[1]Moods Median Test was used to compare medians.
Height adjustments are not required for Lung clearance index (n turnovers) and t_{PTEF}/t_E.
t_{PTEF}/t_E = time to peak tidal expiratory time over total expiratory time
Ree = respiratory resistance at the end of expiration; Xee = respiratory reactance at the end of expiration

Table 4.13: Lung function outcomes by LRTI history: never versus ever diagnosed

The effect of LRTI frequency, timing, and hospitalisation on lung function was further investigated in Tables A.3 and A.4, shown in the appendix. Each row in Tables A.3 and A.4 indicates the adjusted estimates from separate models for each LRTI stratification that included age (months), a dummy variable indicating the different periods (< 3 years or > 3 years), and the full set of predictors: sex (male), preterm (<37 weeks), HIV exposed uninfected versus HIV unexposed, birthweight z-score, wheeze phenotype, LRTI ever, postnatal IPV, maternal asthma/allergy, postnatal smoking, maternal education, asset ownership, household size, and income. Every additional episode of LRTI further increased respiratory rate (1.012 [95% CI 1.009-1.016]) and LCI (1.002 [1.001-1.004]) and reduced tidal volume (0.996 [0.993-0.998]) and compliance (0.987 [0.98-0.99]). If LRTI occurred before age 1 year, there was an additional reduction in tidal volume (0.98 [0.97-0.99]) and an increase

in respiratory rate (1.03 [1.01-1.04]). The effect of hospitalised and ambulatory LRTI was similar; however, this differed for RSV, as RSV LRTI requiring hospitalisation was associated with impaired lung function development compared with an ambulatory RSV LRTI, with lower compliance (0.94 [0.90-0.98]) up to age 5 years. RSV LRTI compared with non-RSV LRTI was associated with a lower t_{PTEF}/t_E (0.97 [0.95-0.99]), consistent with airflow obstruction. Rhinovirus LRTI compared with non-rhinovirus LRTI was not associated with lower lung function in this cohort.

Figure 4.13 displays forest plots which is a useful graphical display of the model output from Tables A.3 and A.4. These plots are especially useful for visually summarising the estimated effects and their uncertainties. In this context, the forest plots depict the model output for several key respiratory measures: resistance, compliance, LCI, and respiratory rate.

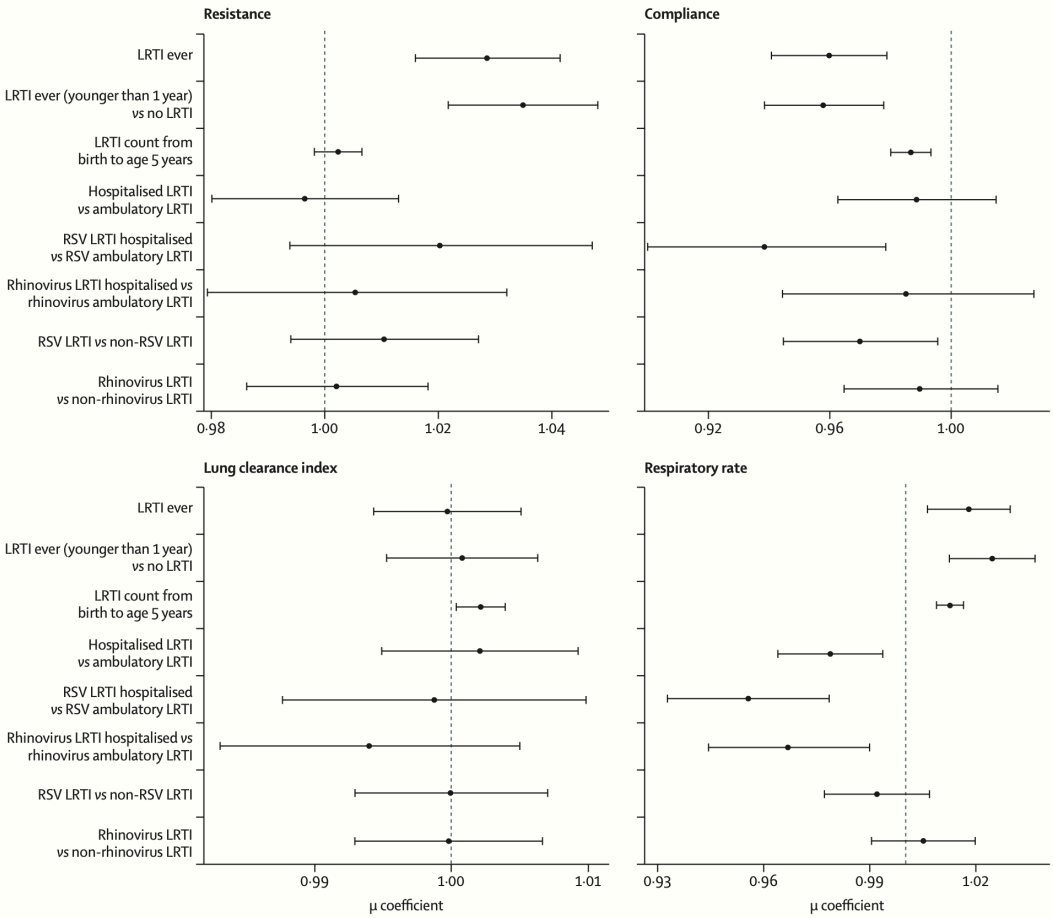


Figure 4.13: Association of specified LRTIs with the development of lung function parameters from birth to 5 years.

The impact of multiple exposures

To highlight for clinicians which children may be most at risk, figures have been included in this section to track lung function in at-risk groups (both with single key risk factors and composite factors) compared to unexposed children.

The impact of multiple exposures over 5 years is shown graphically on resistance and compliance outcomes across various subgroups of interest, Figure 4.14. Resistance and compliance are chosen as they are believed to be sensitive measures which would show clinically clear differences between at-risk and unexposed children, [11, 64]. Figure 4.14 illustrates how multiple exposures appear additive, with children with co-existing preterm, postnatal smoking, and LRTI having the poorest lung development. This illustrates the lasting effects that early life factors can have on health risks throughout a child's life. It underscores the importance of identifying specific subgroups of children who could be prioritised for preventive interventions.

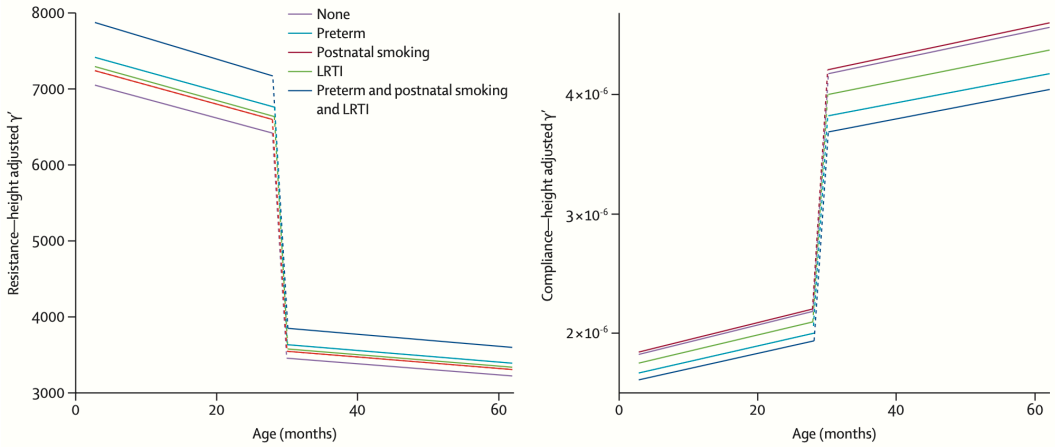


Figure 4.14: Predicted height-adjusted resistance and compliance by exposure groups over time.

4.3.7 Latent profiles of longitudinal lung function outcomes

Aims and background

The goal of this section is to explain the heterogeneous structures within the lung function profiles by exploring different latent class structures and increasing the number of classes. For this, LCMM was chosen over other approaches because it is specifically designed for modelling complex longitudinal data with unobserved heterogeneity. It allows for the identification of latent classes, capturing differences between subgroups and their distinct trajectories while incorporating random effects and covariates.

The beginning stages of this section will be focused on age-related class differences, where only the shape of the age-related trajectory defines the latent classes. After this, a more flexible approach is applied by exploring potential heterogeneity in how certain risk factors affect the latent class structure. Using latent profile identification provides a more nuanced insight into the cohort's heterogeneity.

Single trajectory prediction, like GAMLSS, uses predictors (e.g., LRTI, sex, smoke exposure) to influence a single, average lung function trajectory for the entire cohort. In contrast, latent profile analysis applies similar predictors, but now to classify individuals into different latent classes, each with its own trajectory. This allows for a more nuanced understanding of how covariates affect lung function across heterogeneous subpopulations. The two approaches are linked in that the insights from single trajectory prediction can guide the selection of covariates used in the analysis of latent classes.

Modelling setup

The implemented modelling setup is explained using the outcomes for airwave oscillometry, although the same setup was followed for the MBW and tidal breathing outcomes. The analysis begins with trajectories of latent classes within univariate outcomes for resistance, compliance, Ree, and Xee using LCMM. These four outcomes are categorised as airway oscillometry measurements. Each outcome is represented through LMS z-scores.

To capture the heterogeneity of the data, models of increasing complexity are fitted, including linear, quadratic, and spline models, as done previously when modelling single trajectories. These models are tested to the presence of 1-4 underlying latent classes. The approach is then extended to the multivariate LCMM framework, allowing for the simultaneous modelling of multiple lung function outcomes. The outcome variables measure the same or similar underlying processes, thereby enabling a more comprehensive understanding of respiratory development. Z-scores were modelled which were calculated using the LMS method within the GAMLSS framework. As multiple outcomes are modelled, the LMS z-scores are useful, since all the outcomes will be placed on a normal distribution curve (i.e. placing them on the same scale). Figure 4.15 shows the steps in the approach to this problem. The same approach was employed for tidal breathing and multiple breath washout

measures.

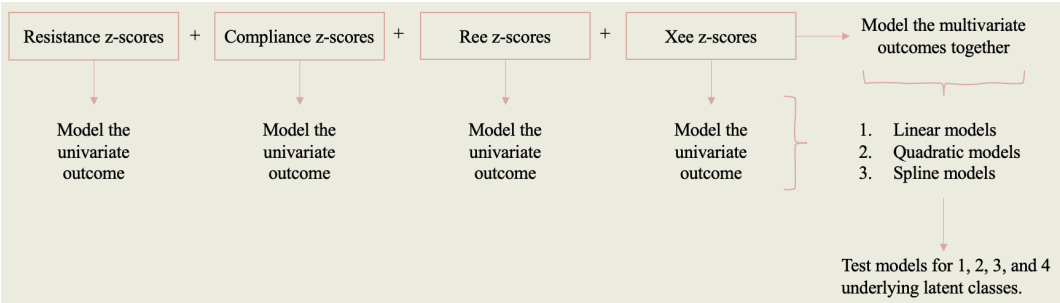


Figure 4.15: Analysis approach to univariate and multivariate longitudinal outcomes for airwave oscillometry

The progression of the models reflects an increasing complexity in modelling the relationship between the response variable and the predictor variable, *Age*. In the first model, only a simple linear relationship is considered, where the response variable is modelled as a constant term (intercept) plus a linear term of *Age*. This assumes a straight-line relationship between *Age* and the response. The second model builds upon the first model by introducing a quadratic term, I(Age^2). This allows for a more flexible relationship between *Age* and the response variable. The inclusion of the quadratic term captures the possibility of a nonlinear pattern, such as a U-shaped or inverted U-shaped relationship. In model 3 (spline), the complexity increases further by using a cubic B-spline basis expansion of *Age* instead of simple linear or quadratic terms. The *splines::bs* function is used to create a smooth, nonlinear representation of the relationship between *Age* and the response variable using basis splines. This allows for even more flexibility in capturing complex, nonlinear relationships that may not be adequately captured by a simple linear or quadratic model.

Overall, the progression of the models represents a move from simpler to more flexible and expressive forms of modelling should enable a better understanding of the underlying patterns and potential nonlinearities in the data.

LCMM results for airwave oscillometry, MBW, and tidal breathing - single longitudinal outcomes

This section presents the latent class structures derived from LCMM analysis for both airwave oscillometry, MBW, and tidal breathing outcomes. Two approaches were used to illustrate the progression of the classes. Firstly, the smoothed longitudinal profiles were plotted using the loess function in R for visualisation purposes, similar to approaches used in published studies, [164, 165, 166]. This technique was employed to provide a clearer and more accurate depiction of the overall trends in the data. Secondly, a Sankey diagram was created for each lung function outcome to highlight the movement of subjects from one class to another.

Spline models were found to consistently produce strong goodness-of-fit statistics, likely due to several factors. First, z-scores, which normalise the response variable relative to the mean, often uncover intricate relationships that may be overlooked when using raw data. Spline models are particularly well-suited to capture these subtle variations in z-scores across different age groups. Additionally, because z-scores emphasise extreme values, splines - through their localised fitting - can model these outliers more accurately, allowing the curve to flexibly adapt to the shape of the trajectories. Unlike more rigid models, such as quadratic forms, spline models allow local regions of the data to behave differently than others (via knots), which is important in z-score data where age-related changes might not be uniform across age. At the same time, splines maintain global smoothness. By using localised fitting, spline models can adapt to nonlinear growth patterns seen in respiratory development, reflecting the reality that airway function may decline or stabilise at different ages for different individuals.

Figure 4.16 displays the identified trajectories of each latent class for each of the univariate airwave oscillometry outcomes using spline models, illustrating the progression as the number of classes increases from 1 to 4. The general pattern observed showed that the transition from a 1-class to a 2-class solution results in the mean profile splitting into two oscillating profiles. Introducing a third class identifies a group that remains stable over time, which may reflect a group with better-managed or less severe symptoms. With the addition of a fourth class, the patterns become more variable and are influenced by a small number of observations. Figure 4.16 is interpreted along with Table 4.14. The goodness-of-fit statistics of the different spline models are shown in Table 4.14. The BIC suggests a 2-class model is optimal. In the 2-class solution, all class sizes are greater than 10%, except for Xee.

Similarly, for MBW and tidal breathing outcomes, Figure 4.18 shows that under the 1 class solution, the trajectories for all outcomes are centred around a single curve. Introducing a 2-class solution separates the population into a dominant majority class and a smaller, deviating class. This separation could reflect a subgroup with more rapid disease progression or greater susceptibility to airway obstruction. As the number of classes increases to 3 or 4, patterns become more variable, often driven by small sample sizes. For example, in the 4-class solution for LCI, smaller classes consist of very few observations, raising concerns about whether these represent true, generalisable patterns or artefacts of noise.

Sankey diagrams in Figures 4.17, with accompanying cross-tabulations, offer additional insights into the clinical progression of latent classes. As the number of latent classes increases, both compliance and Xee exhibit a prominent majority class that gradually diminishes in size. Consequently, smaller classes emerge, which can be viewed as classes comprising outlier points. The Sankey diagram depicting resistance and Ree reveals that, when moving from a 2-class to a 3-class solution, there is an inclination for relatively large groups of subjects from different classes to reunite within the same class after initially separating from each other. This may indicate that some subjects experience periods of instability followed by a return to baseline, a pattern known in lung diseases like asthma.

Further univariate results are shown in the Appendix, Figures A.10 and A.11 illustrate the univariate airwave oscillometry outcomes using linear and quadratic models. Their model fit statistics are shown in Appendix Tables A.7 and A.8. As model complexity increases (linear to polynomial to splines), fewer cases of empty classes where no solution was obtained were observed.

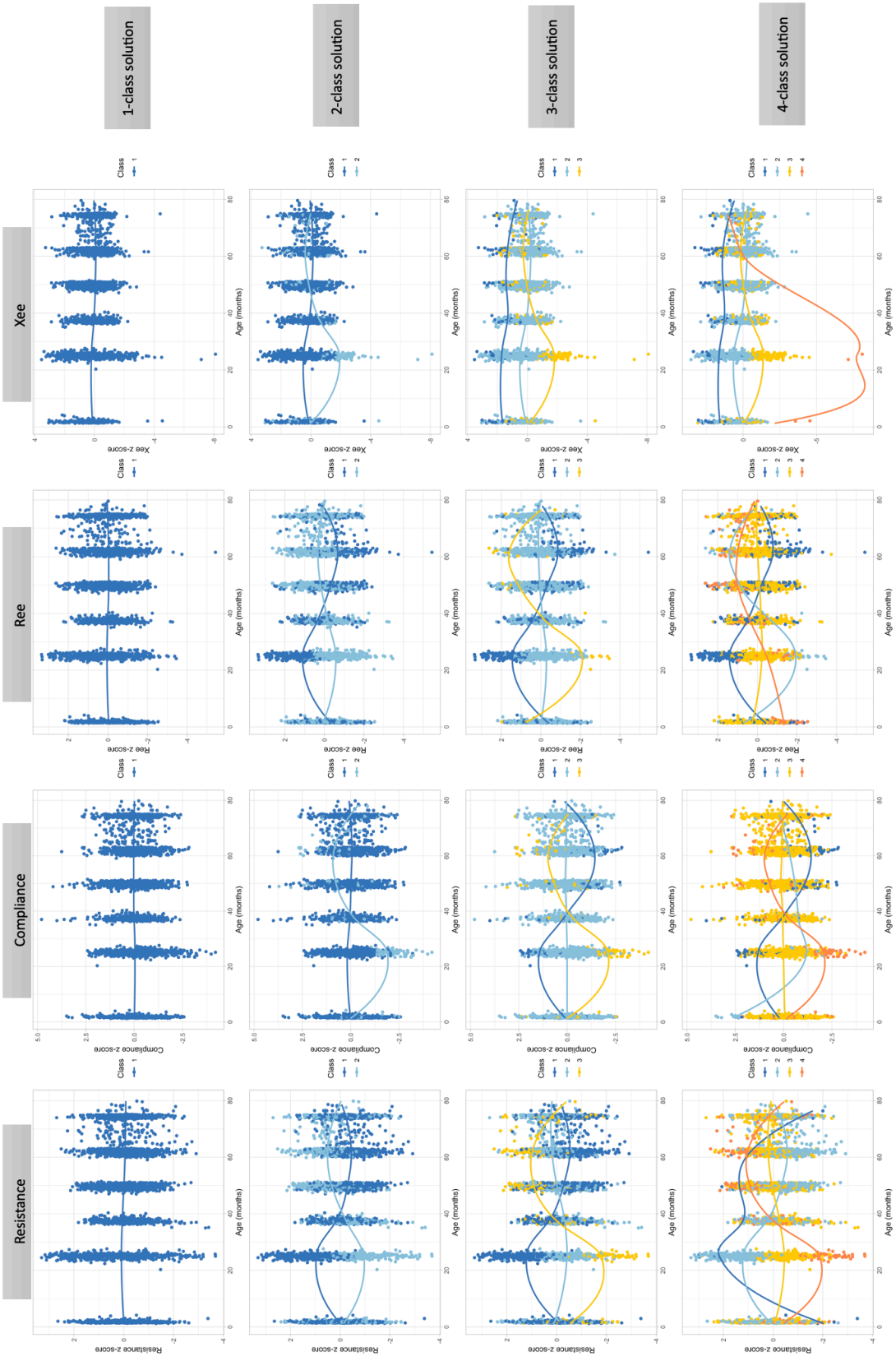
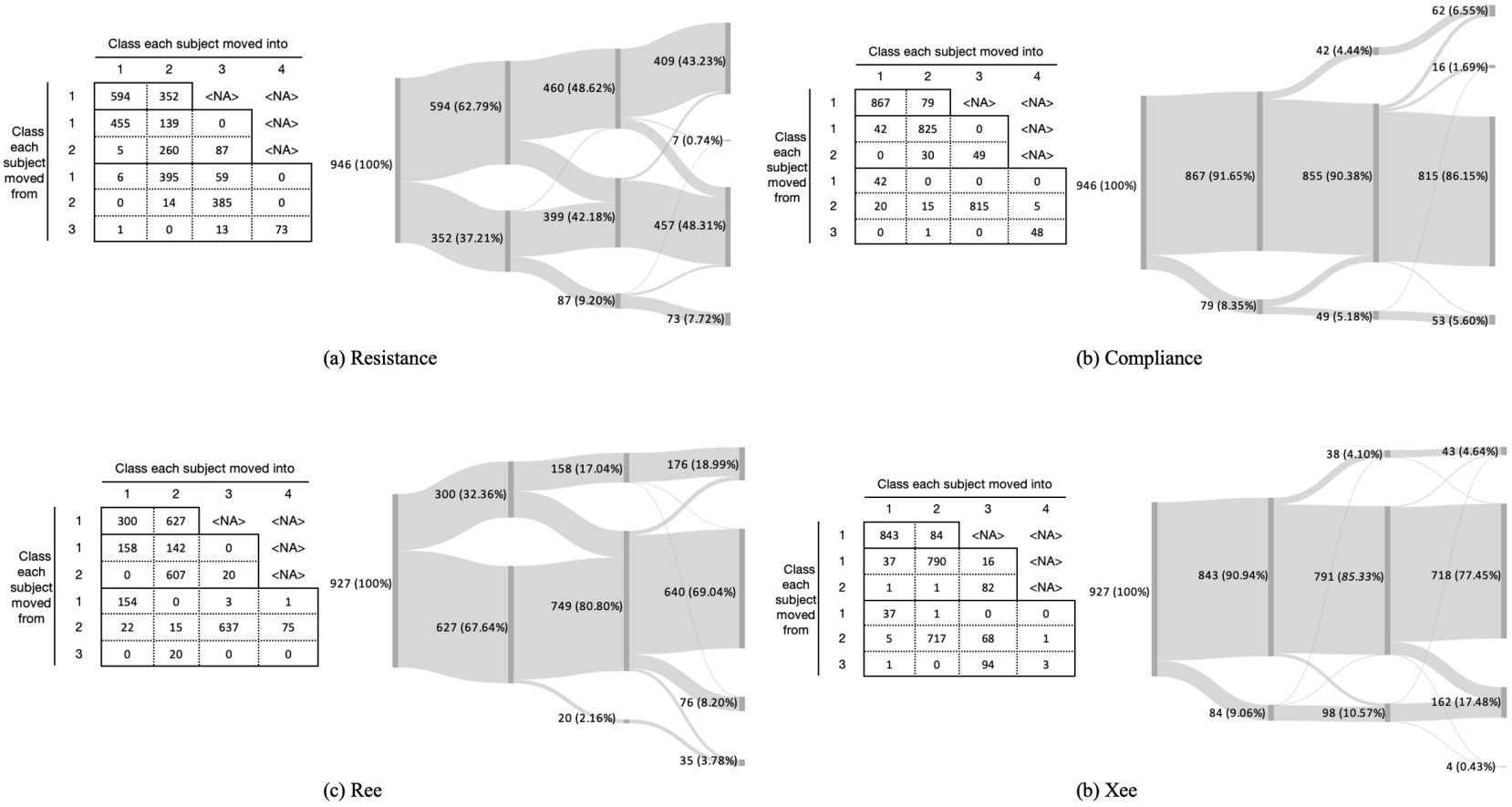


Figure 4.16: This figure illustrates the heterogeneous structures within longitudinal univariate airwave oscillometry outcomes using spline models. Points show individual scores, while classes were computed through the loess function in R for visualisation purposes.



(a) Resistance

(b) Compliance

(c) Ree

(b) Xee

Figure 4.17: Cross tabulations and Sankey diagrams illustrates how the heterogeneous structures developed within longitudinal univariate airwave oscillometry outcomes, using spline models, as we increase the number of classes from 1 to 4.

Table 4.14: Summary table: Univariate airwave oscillometry spline models depicted in Figure 4.16

Variable	G	loglik	npm	AIC	BIC	entropy	Class 1 (%)	Class 2 (%)	Class 3 (%)	Class 4 (%)
Resistance	1	-4655.37	8	9326.75	9365.57	1.00	100.00%			
	2	-4551.25	13	9128.50	9191.58	0.40	62.79%	37.21%		
	3	-4543.01	18	9122.03	9209.37	0.39	48.62%	42.18%	9.20%	
	4	-4537.22	23	9120.58	9232.18	0.50	0.74%	43.23%	48.31%	7.72%
Compliance	1	-4677.59	8	9371.18	9409.99	1.00	100.00%			
	2	-4643.38	13	9312.77	9375.85	0.57	91.65%	8.35%		
	3	-4639.19	18	9314.39	9401.73	0.51	4.44%	90.38%	5.18%	
	4	-4627.54	23	9301.08	9412.68	0.55	6.55%	1.69%	86.15%	5.60%
Ree	1	-4206.33	8	8428.66	8467.31	1.00	100.00%			
	2	-4151.70	13	8329.41	8392.23	0.33	32.36%	67.64%		
	3	-4142.81	18	8321.62	8408.60	0.51	17.04%	80.80%	2.16%	
	4	-4139.80	23	8325.60	8436.74	0.39	18.99%	3.78%	69.04%	8.20 %
Xee	1	-4156.694	8	8329.388	8368.044	1.00	100.00%			
	2	-4086.726	13	8199.451	8262.267	0.63	90.94%	9.06%		
	3	-4073.327	18	8182.653	8269.628	0.65	4.10%	85.33%	10.57%	
	4	-4021.006	23	8088.012	8199.147	0.66	4.64%	77.45%	17.48%	0.43%

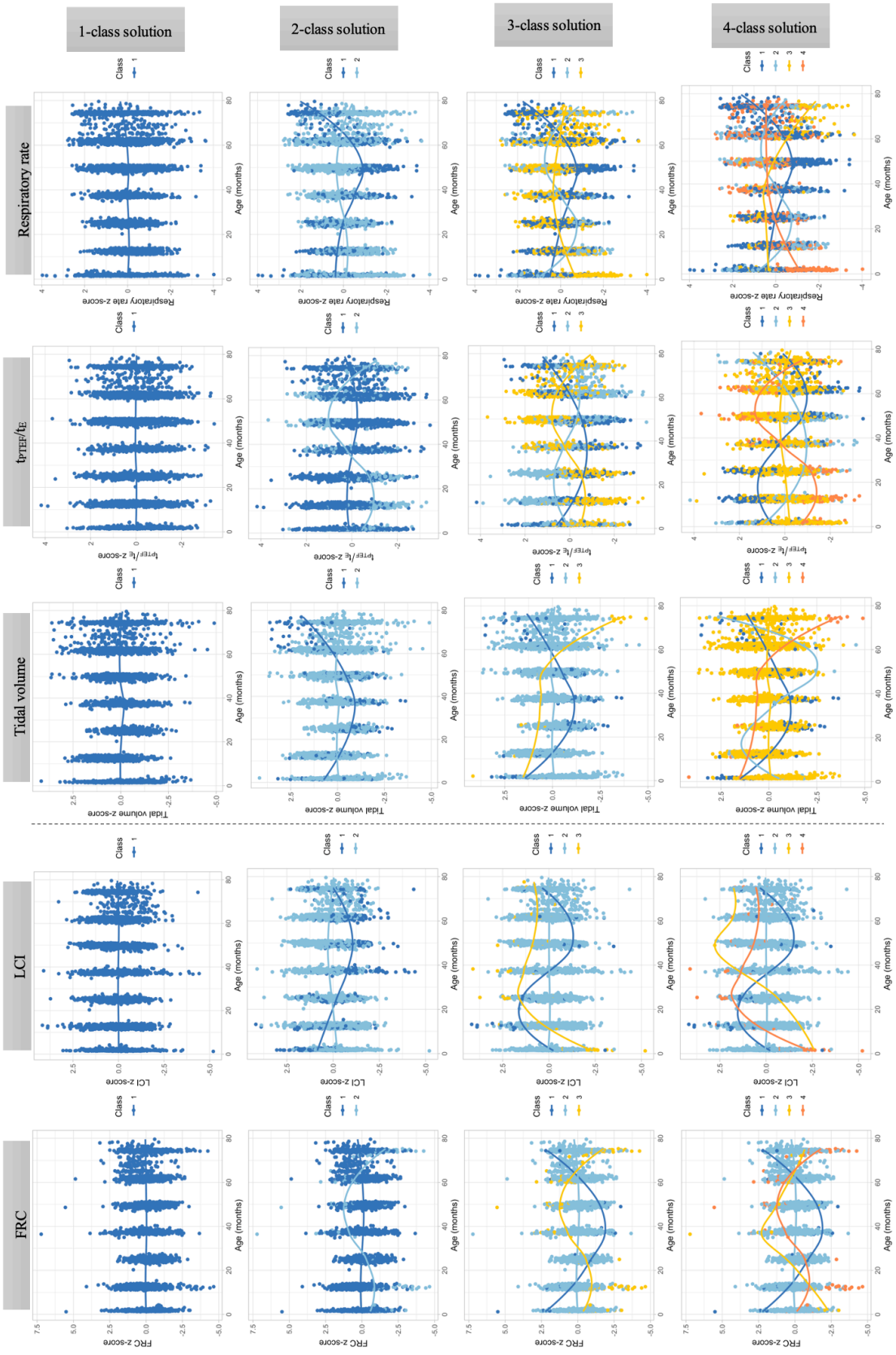


Figure 4.18: This figure illustrates the heterogeneous structures within longitudinal univariate MBW (left) and tidal breathing (right) outcomes using spline models. Points show individual scores, while classes were computed through the loess function in R for visualisation purposes.

Table 4.15: Summary table: Univariate MBW and tidal breathing spline models depicted in Figure 4.18

Variable	G	loglik	npm	AIC	BIC	entropy	Class 1 (%)	Class 2 (%)	Class 3 (%)	Class 4 (%)
FRC	1	-5407.42	8	10830.85	10869.78	1.00	100.00			
	2	-5385.03	13	10796.06	10859.32	0.61	94.68	5.32		
	3	-5376.59	18	10789.18	10876.76	0.73	0.94	93.95	5.11	
	4	-5371.50	23	10788.99	10900.91	0.78	0.94	94.16	0.42	4.48
LCI	1	-5436.26	8	10888.52	10927.44	1.00	100.00			
	2	-5431.24	13	10888.47	10951.73	0.19	17.21	82.79		
	3	-5415.32	18	10866.64	10954.23	0.79	1.56	96.77	1.67	
	4	-5412.94	23	10871.88	10983.80	0.83	1.56	96.87	0.10	1.46
Tidal volume	1	-5932.58	8	11881.16	11920.13	1.00	100.00			
	2	-5924.32	13	11874.64	11937.96	0.31	9.23	90.77		
	3	-5912.39	18	11860.78	11948.46	0.62	5.71	93.36	0.93	
	4	-5901.41	23	11848.82	11960.86	0.69	6.33	0.62	92.12	0.93
t_{PTEF}/t_E	1	-5989.29	8	11994.58	12033.55	1.00	100.00			
	2	-5956.84	13	11939.67	12003.00	0.38	81.22	18.78		
	3	-5940.58	18	11917.15	12004.83	0.29	21.06	43.67	35.27	
	4	-5935.06	23	11916.11	12028.15	0.38	13.90	11.20	68.57	6.33
Respiratory rate	1	-5924.31	8	11864.61	11903.58	1.00	100.00			
	2	-5898.16	13	11822.33	11885.65	0.25	26.66	73.34		
	3	-5878.31	18	11792.63	11880.31	0.30	35.68	14.63	49.69	
	4	-5874.77	23	11795.53	11907.57	0.29	40.77	17.53	10.37	31.33

LCMM results for airwave oscillometry, MBW, and tidal breathing - multiple longitudinal outcomes simultaneously

The lung function outcomes were categorised into three distinct groups based on the testing methods employed during the modelling of multivariate outcomes. Airwave oscillometry (resistance, compliance, Ree, and Xee) measures were modelled together, tidal breathing (tidal volume, respiratory rate, t_{PTEF}/t_E) measures were modelled together, and multiple breath washout (FRC, LCI) measures were modelled together.

The identified trajectories of the multivariate airwave oscillometry outcomes using spline models are presented in Figure 4.19. These figures reflect how combining multiple outcomes, rather than modelling them individually, may uncover more clinically meaningful profiles of lung function development. Similarly, Figure 4.20 provides the results for MBW and tidal breathing outcomes, although no 4-class solution was obtained for tidal breathing. The multivariate models show class-specific trends that span all lung function outcomes in each group, illustrating how latent classes differentiate children with shared profiles of lung function development.

The progression from univariate to multivariate profiles shows how modelling the lung function outcomes together leads to possibly more meaningful profiles. As opposed to the univariate cases where we often see smaller classes of outlying points as the number of classes we investigate increases. These smaller classes may represent atypical or extreme subgroups of children. However, in the multivariate setting, groups of children were identified who follow broadly similar profiles across outcomes.

In a univariate model, only one outcome is being analysed, which limits the amount of information available to identify latent classes. Many lung function measures, such as resistance and compliance, or LCI and FRC, are clinically linked. Univariate models cannot capture these relationships, whereas multivariate models explicitly account for these correlations. A reason for this is, that under the multivariate framework, a single latent process is assumed to underlie all the longitudinal outcomes, which means that each outcome is modelled as being driven by the same unobservable trajectory. This latent process allows the model to capture the correlation between the different outcomes by linking all outcomes to a common source of variation. This means the latent classes are not formed independently for each outcome but reflect joint developmental patterns across multiple outcomes. Modelling related outcomes together in a multivariate framework can help to stabilise class solutions. Further, an unstructured variance-covariance matrix for the random effects was used, which allowed for the estimation of covariances between the random effects of different outcomes. Where no predefined structure is imposed on the relationships between random effects; instead, the model allows the data to inform how the random effects for different outcomes relate to each other.

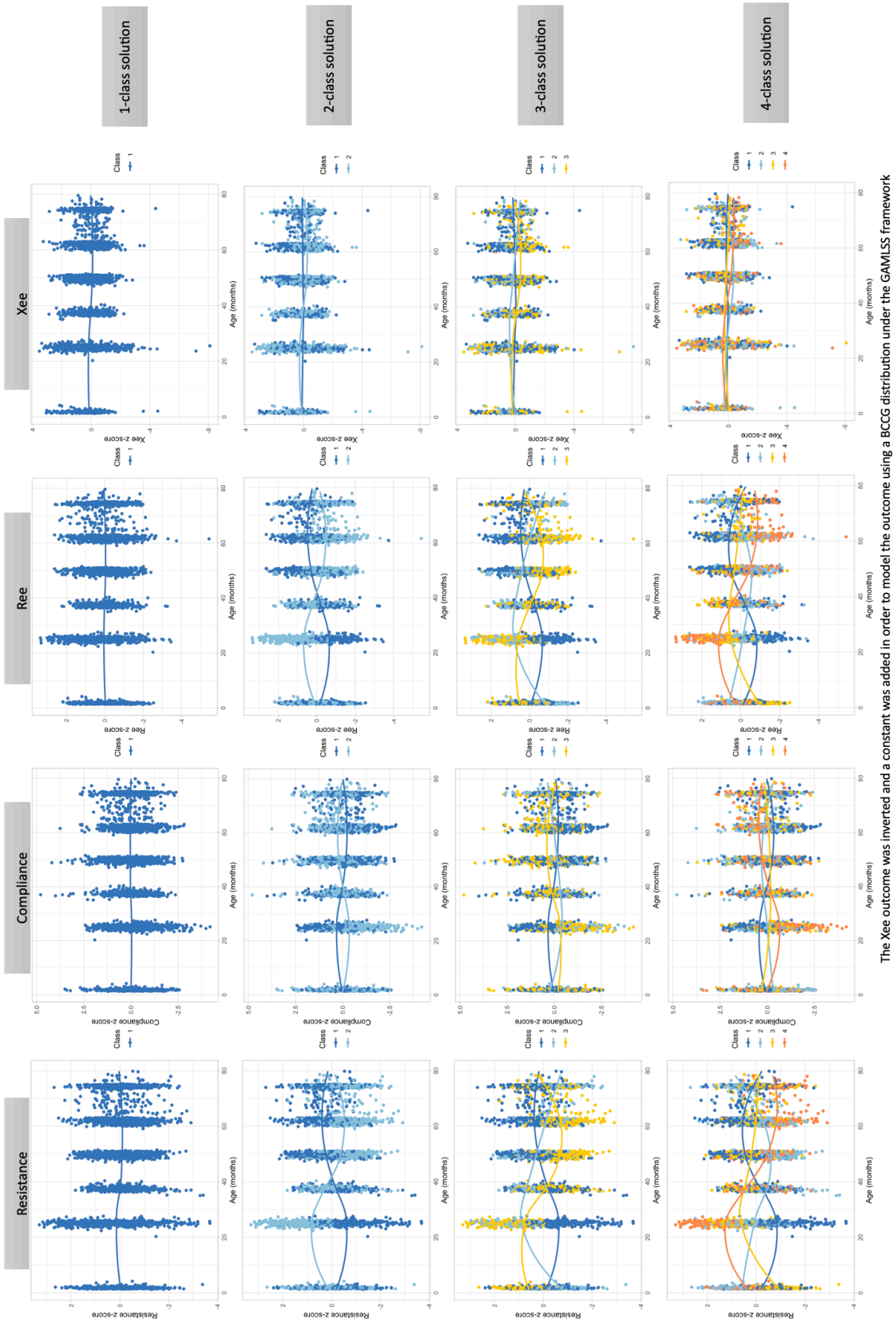


Figure 4.19: This figure illustrates the heterogeneous structures within longitudinal multivariate airwave oscillometry outcomes using spline models. Points show individual scores, while classes were computed through the loess function in R for visualisation purposes.

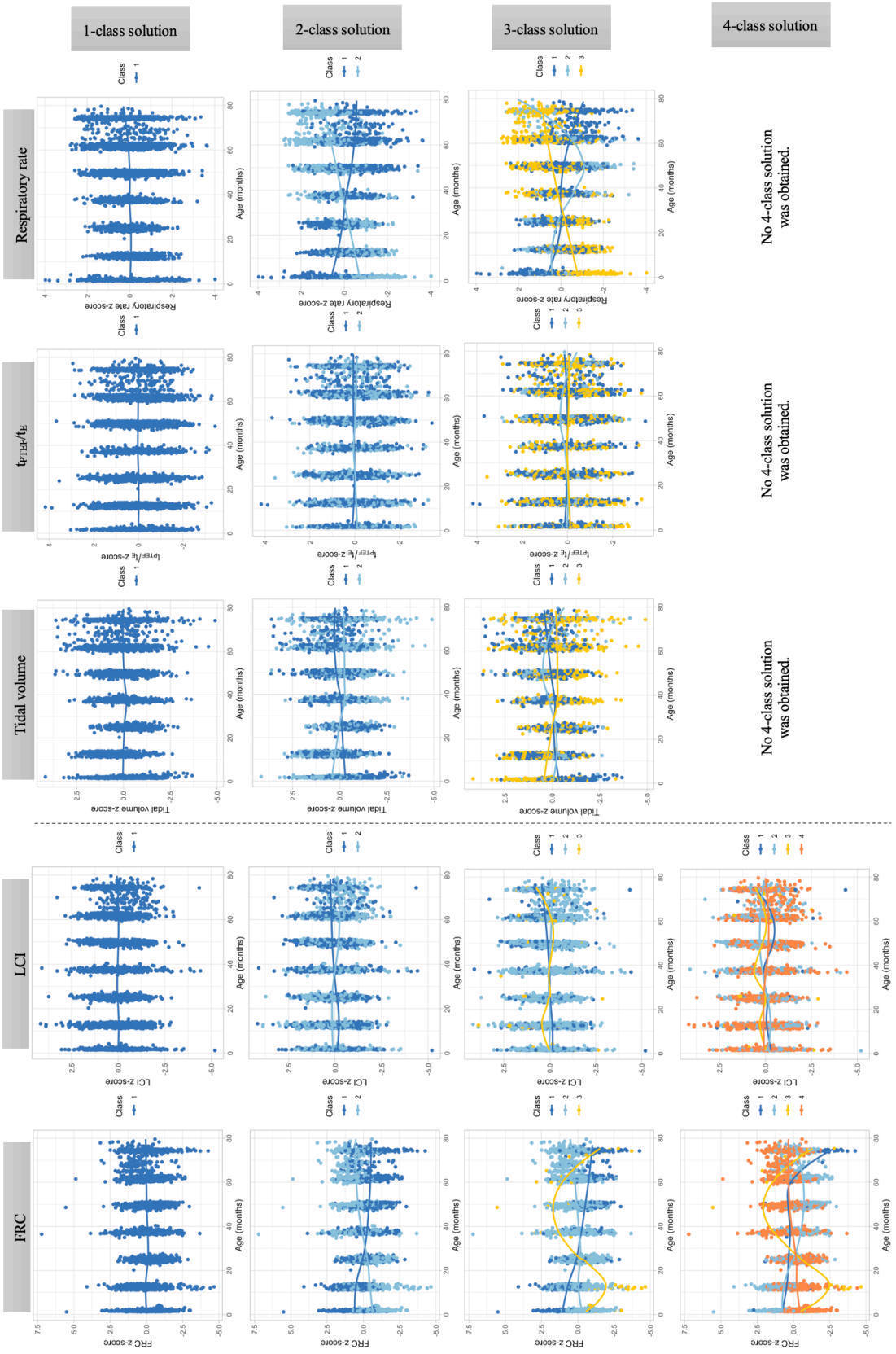


Figure 4.20: This figure illustrates the heterogeneous structures within longitudinal multivariate MBW and tidal breathing outcomes using spline models. Points show individual scores, while classes were computed through the loess function in R for visualisation purposes.

Table 4.16: Summary table: Multivariate airwave oscillometry, MBW, and tidal breathing spline models

G	loglik	npm	AIC	BIC	entropy	Class 1 (%)	Class 2 (%)	Class 3 (%)	Class 4 (%)
					Airwave oscillometry				
1	-17741.22	17	35516.44	35598.93	1.00	100.00%			
2	-28891.47	22	57826.94	57933.69	0.57	46.51%	53.49%		
3	-28816.25	27	57686.50	57817.51	0.57	46.19%	19.56%	34.25%	
4	-28706.71	32	57477.42	57632.69	0.56	31.40%	30.44%	19.66%	18.50%
					MBW				
1	-10872.47	11	21766.95	21820.47	1.00	100.00%			
2	-17893.48	16	35818.95	35896.81	0.21	50.16%	49.84%		
3	-17875.43	21	35792.85	35895.04	0.49	19.19%	78.21%	2.61%	
4	-17861.69	26	35775.38	35901.89	0.52	2.09%	32.12%	1.36%	64.44%
					Tidal breathing				
1	-18099.78	14	36227.55	36295.75	1.00	100.00%			
2	-29747.00	19	59532.01	59624.56	0.39	53.32%	46.68%		
3	-29721.89	24	59491.77	59608.68	0.40	45.75%	11.00%	43.26%	

4.3.8 A feature-based approach to cluster longitudinal lung function

Background

The LCMM results indicated that for certain lung function outcomes, such as compliance and Xee, there was minimal deviation from a single trajectory. In these cases, a single profile would have been sufficient. This section aims to set up a feature-based clustering result which shows at least two clear, well-defined trajectories of appropriate size. Figure 4.21 illustrates the scenario that this feature-based approach aims to address. This is not to say that the LCMM solution which contains a large majority class is incorrect, but rather to provide an alternative approach to the current scenario which could in some cases produce better context-dependent "correctness". In other words, whether the solution aligns more strongly with the specific goals of the analysis or how closely it matches insights from domain knowledge.

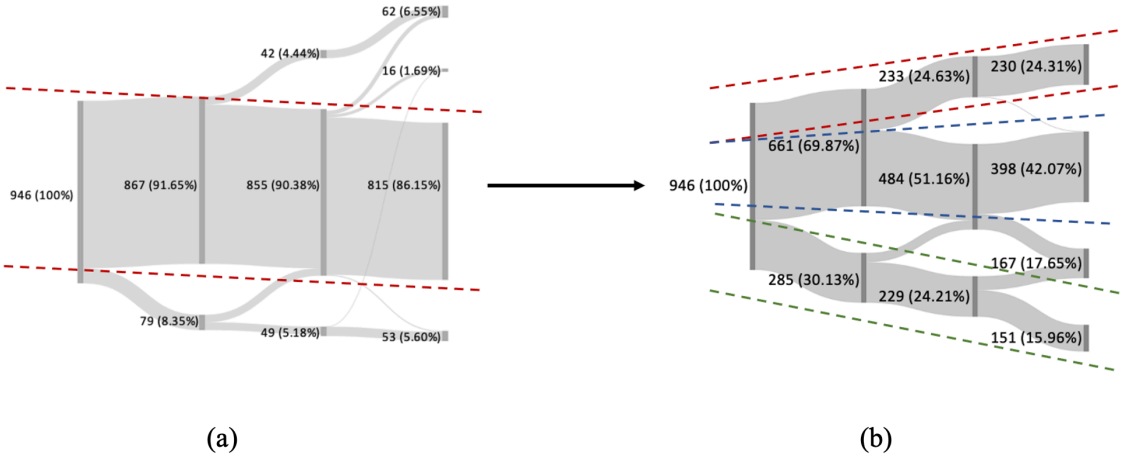


Figure 4.21: Illustration of trajectory clustering in lung function outcomes using compliance as an example. (a) The LCMM analysis from Figure 4.17 shows minimal deviation from a single trajectory. (b) In contrast, the feature-based approach aims to produce at least two distinct, well-defined trajectories of sufficient size.

This approach describes a two-stage method combining GAMLSS modelling for individual trajectories followed by PAM clustering based on estimated parameters. The approach does share similarities with existing statistical methodologies but also has some novel aspects. While random effects from mixed models (e.g., LCMM, LMMs, or GMMs) have been used for clustering, using GAMLSS for feature extraction in a two-stage clustering approach is less common. Therefore, this specific combination (GAMLSS + PAM) has not been used for lung function trajectory analysis, it should be seen as a novel application rather than a wholly new method.

This approach is done in two main steps, detailed by Algorithm 3. Step one requires the specification

of a GAMLSS model in which one mean profile with random errors is modelled. Given a simple linear model, a random intercept and random slope are assigned for each child. Step two requires calculating a distance matrix from the estimates obtained from the fitted model in step one. The distance matrix is then used as input into the PAM algorithm.

When used within the GAMLSS framework, it is necessary to decide on the appropriate distributional family. In this specific scenario, $y = (-\infty, \infty)$, $\mu = (-\infty, \infty)$, and $\sigma > 0$. Normal distributions or t-family distributions, which generalise the normal distribution, can be fitted to the data.

In step one, a decision is required around the appropriate distributional family within the GAMLSS framework, the model structure for each parameter of the selected distributional family, and the random effects structure that is to be used. The important output from step one includes the random errors, model fit statistics, and residual plots. In the second step, an appropriate distance matrix needs to be selected, along with a selection criteria (the ASW), and the number of clusters which are to be compared. The important output from step two includes the ASW criteria values, class sizes, and longitudinal plots.

Algorithm 3 Two-step approach for lung function clustering

1: **Step 1: Modelling Approach**
2: Based on the model type (linear, quadratic, or spline), fit an appropriate GAMLSS model using either the t-family or normal distribution, where one mean profile with child-specific random effects is modelled.
   - Linear Model: Fit a GAMLSS model for the lung function outcome over age with random intercept and slope for each child.
   - Quadratic Model: Add a quadratic term for age, including random intercept and slope effects.
   - Spline Model: Use cubic splines to model nonlinear age effects.
3: Evaluate and select the most appropriate model and distribution using the AIC value.
4: **Step 2: Clustering Using PAM**
5: Extract random effects from the fitted model selected in Step 1 for each child.
6: Use these random effects to perform clustering via the PAM algorithm, using an appropriate distance matrix as input.
7: Iterate over possible values of k (number of clusters) from 1 to 4 to generate different clustering solutions and store the ASW (Average Silhouette Width) value.
8: Compare clustering results and report the distribution of children across clusters.

Comparisons and differences can be drawn between the LCMM approach and the two-stage feature-based approach. LCMM and GAMLSS are two different statistical modelling approaches with different goals. LCMM is primarily used to model latent classes within longitudinal data. GAMLSS,

on the other hand, is a flexible framework used for modelling the distributional properties of response variables. All four moments of a distribution can be explicitly modelled by functions of explanatory variables using GAMLSS, while this is not explicitly true for LCMM. GAMLSS does not explicitly model latent classes or group-specific variations, while LCMM explicitly models the existence of unobserved subpopulations. Therefore, to obtain similar classes of subpopulations, a two-stage approach is set up using the GAMLSS framework where the identification of subpopulations is done in a second clustering step using the PAM algorithm based on features from the GAMLSS model.

The main difference between the two approaches is that the clustering is done independently from the modelling in the two-stage feature-based approach using GAMLSS and PAM.

Both LCMM and GAMLSS can loosely be seen as frameworks for fitting regression-type models. The clustering outcomes are compared between results where the outcome was modelled using identical model formulations.

Results

In this section, the initial results and output are displayed for the compliance and Xee outcomes. In the first stage, different models (linear, quadratic, and spline) were fitted to the data using both the t-family (TF) and normal (NO) distributions. The AIC results for compliance, detailed in Table 4.17, highlight the performance of each model type and distributional family combination within the GAMLSS framework. Notably, the TF distribution consistently outperformed the NO distribution across all model types, with the spline model paired with the TF distribution achieving the best fit (AIC = 9313.18).

For the Xee outcome, a similar trend emerged. As shown in Table 4.19, the spline model with the TF distribution achieved the lowest AIC (7984.75), outperforming both the linear and quadratic models.

In the second stage, focused on clustering, Table 4.17 shows the ASW for different class sizes for compliance and was used to assess cluster quality. The Euclidean distance matrix was used with the PAM algorithm, which was found to produce better separation between classes, according to the ASW, when compared to the Manhattan distance. The spline model provided the best fit to the compliance data based on AIC, particularly when using the t-family distribution. However, the linear and quadratic models offered competitive performance in terms of AIC and clustering quality. The 2-class solutions yielded the highest ASW values across all models, suggesting that the compliance data may naturally fall into two distinct groups.

For the Xee outcome, the 2-class solution also showed the highest ASW values (ASW = 0.558 for TF-linear and TF-quadratic models), indicating a clear separation of two clusters across different model types, Table 4.19. However, in the spline model, the ASW values slightly dropped (0.517), but the cluster sizes remained relatively robust.

Key insights from the analysis include the consistent superiority of the t-family distribution in terms of both AIC and ASW across all model types. While models using the normal distribution generally exhibited higher AIC and lower ASW values, they occasionally provided competitive or superior clustering performance when a greater number of clusters was desired, such as in the 4-class solutions for quadratic and linear models for the compliance outcome.

Comparing the two-stage feature-based approach with the LCMM approach

The 2-class solutions yielded the highest ASW values across all models for compliance, offering a contrast to the LCMM results shown in Table 4.14, whereas LCMM favoured a more intricate 4-class model with one dominant class. While the two-stage feature-based approach simplifies the structure to 2 classes, offering stronger cluster coherence. This suggests that while LCMM identifies more latent classes, the two-stage method selects a simpler, more clearly defined grouping. This contrast is further highlighted in Table 4.18, summarising class sizes for scenarios 1, 3, and 5 from Table 4.17. The two-stage feature-based approach avoids clustering solutions which contain small speculative classes for compliance. In contrast, the earlier LCMM results from Table 4.14 showed that LCMM was susceptible to producing clustering solutions with small class sizes for compliance.

A similar comparison can be made for the Xee outcome, where the 2-class solution yielded the highest ASW values. These findings can be contrasted with the LCMM results for compliance shown in Table 4.14. Whereas, LCMM favours a 4-class model, with one dominant class and a speculative fourth class of small size.

The benefit of the two-stage feature-based solution includes the ability to increase the number of classes without ending up with small classes of outliers. The tests further show that we can obtain clustering results which show at least two clear, well-defined trajectories of appropriate size for compliance and Xee.

For both the compliance and Xee outcomes, Figures 4.23 and 4.25, the two-stage approach tends to maintain relatively stable memberships within each class, with limited movement of subjects between classes as the number of clusters grows. This results in a more rigid class assignment that reflects the initial classification. Overall, the two-stage approach emphasises stable classes with minimal reclassification.

Table 4.17: Average silhouette width (ASW) values obtained from step 2 using the PAM algorithm for linear, quadratic, and spline models from step 1 in the two-stage feature-based approach for compliance.

			Stage 1			
	Scenario 1	**Scenario 2**	**Scenario 3**	**Scenario 4**	**Scenario 5**	**Scenario 6**
Model type	Linear	Linear	Quadratic	Quadratic	Spline	Spline
Distribution	TF()	NO()	TF()	NO()	TF()	NO()
AIC	9313.719	9355.418	9314.421	9357.221	9313.182	9356.097
			Stage 2			
1 Class	0.000	0.000	0.000	0.000	0.000	0.000
2 Classes	0.562	0.554	0.561	0.540	0.526	0.340
3 Classes	0.522	0.512	0.519	0.491	0.543	0.259
4 Classes	0.497	0.527	0.495	0.499	0.484	0.275

The euclidean distance matrix was used with the PAM algorithm in stage 2.

Linear	Class (%)			
	Class 1	Class 2	Class 3	Class 4
2 class solution	52.39%	47.61%	/	/
3 class solution	27.84%	37.62%	34.54%	/
4 class solution	22.95%	27.84%	25.72%	23.49%
Quadratic	Class (%)			
	Class 1	Class 2	Class 3	Class 4
2 class solution	51.22%	48.78%	/	/
3 class solution	28.05%	36.87%	35.07%	/
4 class solution	22.85%	28.05%	25.50%	23.59%
Spline	Class (%)			
	Class 1	Class 2	Class 3	Class 4
2 class solution	69.87%	30.13%	/	/
3 class solution	24.63%	24.21%	51.16%	/
4 class solution	24.32%	15.96%	42.07%	17.65%

Table 4.18: Summary table, for compliance, of the class sizes for scenario 1, 3, and 5 in Table 4.17.

Table 4.19: Average silhouette width (ASW) values obtained from step 2 using the PAM algorithm for linear, quadratic, and spline models from step 1 in the two-stage feature-based approach for Xee.

			Stage 1			
	Scenario 1	**Scenario 2**	**Scenario 3**	**Scenario 4**	**Scenario 5**	**Scenario 6**
Model type	Linear	Linear	Quadratic	Quadratic	Spline	Spline
Distribution	TF()	NO()	TF()	NO()	TF()	NO()
AIC	8089.37	8288.56	8091.83	8290.41	7984.75	8114.56
			Stage 2			
1 Class	0.000	0.000	0.000	0.000	0.000	0.000
2 Classes	0.558	0.545	0.558	0.545	0.517	0.529
3 Classes	0.524	0.504	0.525	0.504	0.507	0.546
4 Classes	0.521	0.506	0.518	0.507	0.449	0.484

The euclidean distance matrix was used with the PAM algorithm in stage 2.

Linear	Class (%)			
	Class 1	Class 2	Class 3	Class 4
2 class solution	42.83%	57.17%	/	/
3 class solution	29.67%	38.51%	31.82%	/
4 class solution	24.06%	36.35%	10.25%	29.34%
Quadratic	Class (%)			
	Class 1	Class 2	Class 3	Class 4
2 class solution	44.34%	55.66%	/	/
3 class solution	29.88%	38.83%	31.28%	/
4 class solution	26.87%	34.84%	11.43%	26.86%
Spline	Class (%)			
	Class 1	Class 2	Class 3	Class 4
2 class solution	61.60%	38.40%	/	/
3 class solution	44.55%	27.62%	27.83%	/
4 class solution	33.55%	22.76%	18.45%	25.24%

Table 4.20: Summary table, for Xee, of the class sizes for scenario 1, 3, and 5 in Table 4.19.

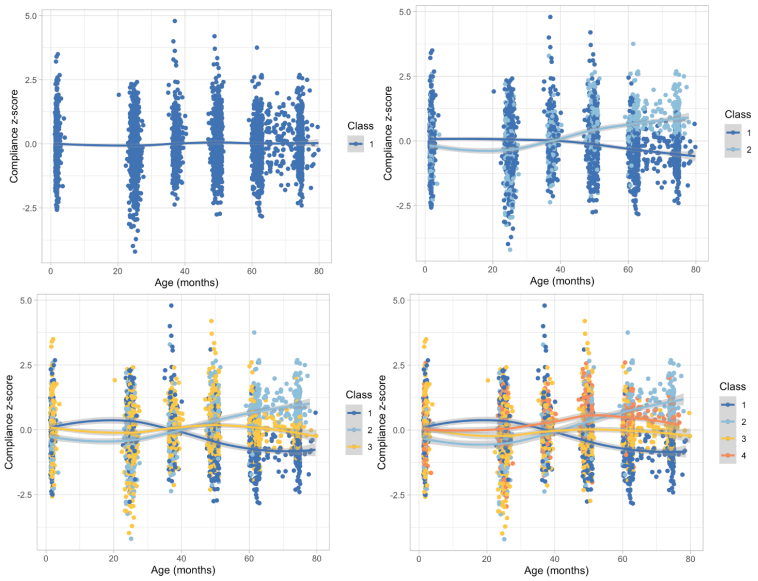


Figure 4.22: This figure illustrates the heterogeneous structures within compliance z-scores using the two-stage feature-based approach with spline models. Points show individual scores, while classes were computed through a loess function for visualisation purposes.

Class each subject moved into

Class each subject moved from	1	2	3	4
1	661	285	<NA>	<NA>
1	233	0	428	<NA>
2	0	229	56	<NA>
1	230	0	3	0
2	0	151	0	78
3	0	0	395	89

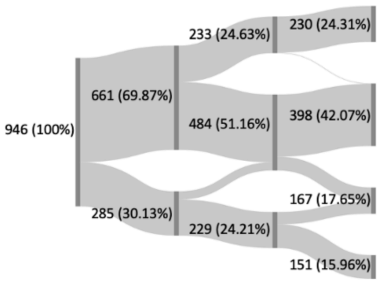


Figure 4.23: Cross tabulations and Sankey diagrams illustrates how the heterogeneous structures developed within the compliance outcome, using a feature-based approach, as we increase the number of classes from 1 to 4.

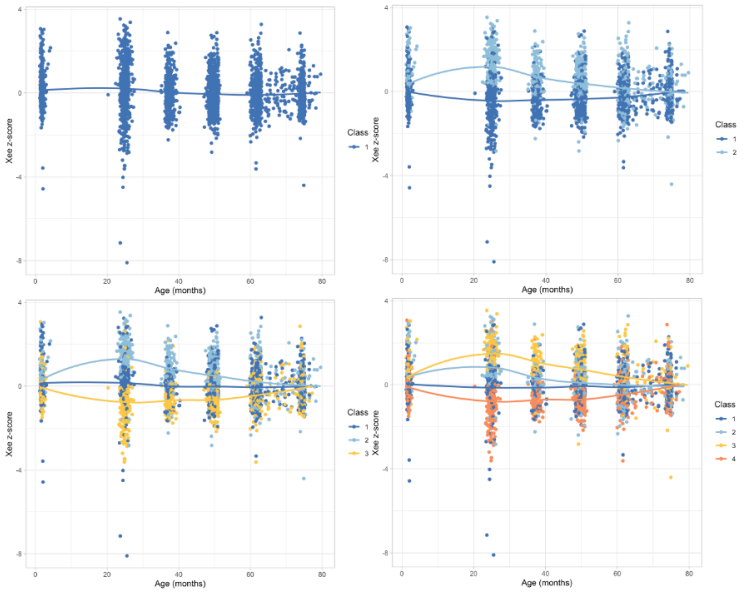


Figure 4.24: This figure illustrates the heterogeneous structures within Xee z-scores using the two-stage feature-based approach with spline models. Points show individual scores, while classes were computed through a loess function for visualisation purposes.

Class each subject moved from	Class each subject moved into: 1	2	3	4
1	571	356	<NA>	<NA>
1	313	0	258	<NA>
2	100	256	0	<NA>
1	287	126	0	0
2	0	85	171	0
3	24	0	0	234

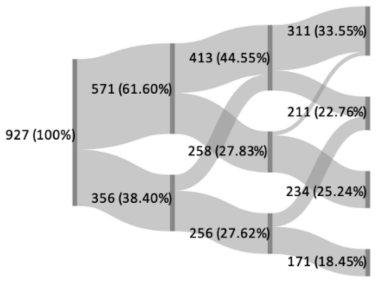


Figure 4.25: Cross tabulations and Sankey diagrams illustrates how the heterogeneous structures developed within the Xee outcome, using a feature-based approach, as we increase the number of classes from 1 to 4.

4.3.9 Early life risk factors and latent classes

This section expands on the earlier analysis in this chapter, which focused on predicting a single trajectory. It draws from those findings to explore and identify the factors influencing distinct lung function trajectories.

There are two primary approaches that can be utilised within the LCMM framework. The choice between these two approaches depends on the desired method for modelling the variation between latent classes and the covariates expected to influence that variation.

Under the first approach, multiple covariates, such as age, sex, preterm, or LRTI, influence the allocation of subjects into different latent classes. This is controlled through the mixture component of the LCMM model. This means that the latent class structure depends on more than just age (the splines function), incorporating other risk factors. Under this modelling approach, the effect of LRTI and sex, for example, would be allowed to vary between latent classes.

While under the second approach, only age (through the splines function) is allowed to vary across latent classes. This means the model is more constrained in how it differentiates between classes, as only the shape of the age-related trajectory defines the latent classes. The latent class structure is simpler compared to the first approach, being solely driven by differences in age trajectories. The other covariates still influence the longitudinal trajectory, but they don't contribute to latent class differentiation.

The first approach (where these risk factors are included in the mixture component) is more appropriate if the interest lies in exploring potential heterogeneity in how these covariates affect the latent class structure. This approach allows these risk factors to differentiate latent classes, which can help explain why subjects are grouped into certain classes based on their risk profiles. By allowing these variables to influence both class membership and the outcome (lung function trajectories), this model captures the heterogeneity in class allocation more fully, reflecting the potential impact of multiple factors beyond just age.

Including risk factors like LRTI, sex, preterm birth, wheezing phenotypes, and birthweight in the mixture component follows clinical sensibility. These factors are known to influence respiratory health and lung function development in children. Allowing them to affect latent class membership recognises that different groups of children may follow distinct lung function trajectories due to underlying biological and environmental factors.

Results for airwave oscillometry

To analyse z-score resistance, compliance, Ree, and Xee patterns, latent class models were fit with 1 to 4 classes using longitudinal data. The earlier modelling results presented in Table 4.10 demonstrated several strong associations between early life risk factors and lung function measures. These findings were used to inform the selection of risk factors for the LCMM model applied to predict the z-score trajectories of these outcomes.

To determine which number of classes produced the best fit across the four outcomes (resistance, Ree, Xee, compliance), the models were compared based on their model summary statistics in Table 4.21. A 2-class model offers the best balance between model complexity and fit. The 2-class model strikes an optimal balance, where adding more classes (e.g., 3 or 4) might lead to diminishing returns in terms of fit while increasing complexity unnecessarily. Both classes tend to be large enough to provide meaningful insights, avoiding the risk of very small, spurious classes that can arise in models with more classes.

Table 4.21: Summary table: LCMM results for z-score resistance, Ree, Xee, and compliance models

Variable	G	loglik	npm	BIC	Class 1 (%)	Class 2 (%)	Class 3 (%)	Class 4 (%)
Resistance	1	-3513.377	10	7092.208	100.00			
	2	-3428.759	15	6955.699	70.26	29.74		
	3	-3421.300	20	6973.507	46.12	39.37	14.51	
	4	-3419.076	25	7001.785	2.44	47.70	38.65	11.21
Ree	1	-3207.809	13	6500.481	100.00			
	2	-3163.438	26	6496.604	31.29	68.71		
	3	-3155.667	39	6565.924	77.05	1.75	21.20	
	4	-3148.614	52	6636.681	58.63	25.44	3.80	12.13
Xee	1	-3207.845	15	6513.609	100.00			
	2	-3163.537	28	6509.856	31.29	68.71		
	3	-3155.554	41	6578.754	20.47	77.78	1.75	
	4	-3152.244	52	6643.942	13.74	79.97	1.90	4.39
Compliance	1	-3485.003	15	7068.186	100.00			
	2	-3448.204	30	7092.769	86.93	13.07		
	3	-3439.937	45	7174.415	11.93	6.47	81.61	
	4	-3429.167	60	7251.055	13.22	5.75	74.28	6.75

Table 4.22 presents the model estimates for resistance, compliance, Ree, and Xee from the LCMM models using splines. The table includes effect sizes with corresponding 95% CI and p-values across two latent classes. Key covariates such as sex, preterm status, birthweight, wheezing phenotypes, LRTI, and maternal factors are included, highlighting their associations with different lung function measures and their latent classes.

Preterm birth was found to be associated with lower compliance in Class 1 and higher values of Ree and Xee in Class 2, coinciding with the modelling results in Table 4.10. Sex (male) appears to be associated with lower Ree in Class 2. In Class 1, LRTI shows a positive association with resistance and a negative association with compliance, indicating stiffer airways. However, there is no strong association with Ree or Xee. In Class 2, LRTI is not strongly associated with resistance or compliance, but it shows strong positive associations with both Ree and Xee, suggesting increased airway resistance and reactance in this group.

	Resistance		Compliance		Ree		Xee	
	Effect size (95% CI)	**P-value**	**Effect size (95% CI)**	**P-value**	**Effect size (95% CI)**	**P-value**	**Effect size (95% CI)**	**P-value**
Intercept Class 1 (not estimated)	0		0		0		0	
Intercept Class 2	-0.092 (-0.485, 0.301)	0.647	0.143 (-0.43, 0.716)	0.625	0.507 (-0.004, 1.018)	0.052	0.433 (0.013, 0.852)	0.043
bs 1 Class 1	3.182 (2.524, 3.84)	0.000	1.156 (0.466, 1.845)	0.001	4.414 (3.132, 5.695)	0.000	4.422 (3.143, 5.702)	0.000
bs 1 Class 2	-3.299 (-4.243, -2.355)	0.000	-4.752 (-6.789, -2.715)	0.000	-2.151 (-3.013, -1.29)	0.000	-2.146 (-3.002, -1.289)	0.000
bs 2 Class 1	-2.729 (-3.301, -2.156)	0.000	-0.883 (-1.565, -0.201)	0.011	-2.911 (-3.832, -1.989)	0.000	-2.913 (-3.831, -1.995)	0.000
bs 2 Class 2	3.004 (2.038, 3.971)	0.000	4.818 (2.826, 6.81)	0.000	1.292 (0.596, 1.988)	0.000	1.288 (0.593, 1.983)	0.000
bs 3 Class 1	0.362 (0.088, 0.637)	0.010	0.207 (-0.058, 0.473)	0.126	0.844 (0.323, 1.365)	0.002	0.844 (0.321, 1.367)	0.002
bs 3 Class 2	-0.039 (-0.477, 0.399)	0.860	-0.748 (-1.538, 0.041)	0.063	-0.231 (-0.592, 0.129)	0.208	-0.230 (-0.589, 0.129)	0.210
Child characteristics								
Sex (male) Class 1	0.016 (-0.089, 0.121)	0.766	0.087 (-0.013, 0.187)	0.088	0.016 (-0.14, 0.172)	0.839		
Sex (male) Class 2	-0.065 (-0.219, 0.088)	0.404	-0.064 (-0.286, 0.157)	0.568	-0.039 (-0.161, 0.083)	0.532		
Preterm (<37 weeks) Class 1	0.112 (-0.044, 0.269)	0.160	-0.153 (-0.286, -0.021)	0.024	0.039 (-0.222, 0.299)	0.772	0.035 (-0.226, 0.297)	0.791
Preterm (<37 weeks) Class 2	0.042 (-0.163, 0.247)	0.689	-0.079 (-0.436, 0.279)	0.667	0.168 (-0.002, 0.338)	0.053	0.168 (-0.001, 0.338)	0.051
Birthweight-z-score Class 1	-0.012 (-0.061, 0.037)	0.628	0.030 (-0.014, 0.073)	0.187				
Birthweight-z-score Class 2	-0.014 (-0.085, 0.056)	0.691	0.067 (-0.039, 0.173)	0.217				
Wheezing phenotype								
Early transient (vs Never) Class 1			0.043 (-0.084, 0.17)	0.509	-0.017 (-0.226, 0.192)	0.871	-0.015 (-0.223, 0.192)	0.887
Early transient (vs Never) Class 2			-0.029 (-0.304, 0.247)	0.838	-0.020 (-0.181, 0.14)	0.804	-0.019 (-0.179, 0.141)	0.815
Late onset (vs Never) Class 1			-0.056 (-0.215, 0.103)	0.488	0.198 (-0.066, 0.462)	0.141	0.202 (-0.061, 0.465)	0.132
Late onset (vs Never) Class 2			-0.141 (-0.541, 0.258)	0.489	0.079 (-0.109, 0.267)	0.411	0.081 (-0.108, 0.269)	0.401
Recurrent (vs Never) Class 1			-0.078 (-0.231, 0.076)	0.319	0.196 (-0.050, 0.443)	0.119	0.197 (-0.049, 0.444)	0.116
Recurrent (vs Never) Class 2			-0.209 (-0.544, 0.126)	0.221	0.034 (-0.154, 0.221)	0.725	0.029 (-0.156, 0.215)	0.757
LRTI								
LRTI ever Class 1	0.100 (-0.005, 0.206)	0.062	-0.148 (-0.258, -0.037)	0.009	0.064 (-0.109, 0.238)	0.466	0.064 (-0.109, 0.238)	0.467
LRTI ever Class 2	0.138 (-0.015, 0.290)	0.078	-0.020 (-0.268, 0.227)	0.871	0.204 (0.068, 0.341)	0.003	0.201 (0.066, 0.336)	0.004
Maternal characteristics								
Postnatal IPV Class 1							-0.005 (-0.18, 0.171)	0.959
Postnatal IPV Class 2							-0.03 (-0.165, 0.105)	0.663
Maternal asthma/allergy Class 1					-0.151 (-0.438, 0.136)	0.302	-0.154 (-0.441, 0.134)	0.295
Maternal asthma/allergy Class 2					-0.067 (-0.286, 0.152)	0.549	-0.059 (-0.276, 0.158)	0.595
Postnatal smoking Class 1	0.01 (-0.097, 0.118)	0.849			-0.077 (-0.249, 0.095)	0.379	-0.076 (-0.255, 0.103)	0.407
Postnatal smoking Class 2	-0.009 (-0.164, 0.146)	0.908			0.004 (-0.117, 0.124)	0.954	0.007 (-0.119, 0.133)	0.913
Socio-economic status (SES)								
Household size								
Small-Medium (vs small) Class 1			0.068 (-0.071, 0.207)	0.338				
Small-Medium (vs small) Class 2			0.058 (-0.29, 0.405)	0.746				
Medium-Large (vs small) Class 1			0.083 (-0.058, 0.225)	0.250				
Medium-Large (vs small) Class 2			-0.068 (-0.415, 0.279)	0.702				
Large (vs small) Class 1			0.107 (-0.031, 0.246)	0.128				
Large (vs small) Class 2			0.157 (-0.147, 0.461)	0.311				

Table 4.22: Model estimates for resistance, compliance, Ree, and Xee from LCMM.

Results for MBW and tidal breathing

To analyse z-scores of MBW and tidal breathing measures, latent class models were fit with 1 to 4 classes using longitudinal data. The earlier modelling results presented in Table 4.9 demonstrated several strong associations between early life risk factors and lung function measures. These findings were used to inform the selection of risk factors for the LCMM model.

To determine which number of classes produced the best fit across the five outcomes (FRC, LCI, respiratory rate, tidal volume, and t_{PTEF}/t_E), the models were compared based on their model summary statistics in Table 4.23. The best-fitting models for each variable were the 2-class models, as they provide the lowest BIC values without overcomplicating the model with additional parameters. Adding more classes tends to increase complexity without improving model fit.

Early transient wheeze (vs Never) was found to be associated with higher LCI in Class 1, consistent with the modelling results in Table 4.9. Sex (male) was associated with higher tidal volume in Class 2, suggesting a sex-specific effect on lung volume development in this group. The observed associations between male sex and higher tidal volumes, as well as the relationship between early transient wheezing and LCI, highlight potential biological and phenotypic factors influencing lung development.

Table 4.23: Summary table: LCMM results for z-score FRC, LCI, RR, tidal volume, and ratio models

Variable	G	loglik	npm	BIC	Class 1 (%)	Class 2 (%)	Class 3 (%)	Class 4 (%)
FRC	1	-2593.908	13	5272.353	100.00			
	2	-2577.256	24	5310.579	6.75	93.25		
	3	-2559.693	35	5346.984	6.30	6.45	87.26	
	4	-2552.592	44	5391.306	12.74	1.80	16.64	68.82
LCI	1	-4161.639	11	8395.387	100.00			
	2	-4147.807	20	8426.721	98.72	1.28		
	3	-4139.161	29	8468.427	1.14	96.59	2.28	
	4	-4135.579	36	8507.151	1.28	96.16	2.42	0.14
Respiratory rate	1	-4581.240	18	9280.552	100.00			
	2	-4507.489	34	9238.005	51.56	48.44		
	3	-4487.119	50	9302.219	44.76	8.36	46.88	
	4	-4474.316	66	9381.566	36.97	10.20	12.89	39.94
Tidal volume	1	-4023.165	14	8138.065	100.00			
	2	-4005.038	26	8180.441	25.96	74.04		
	3	-3991.892	38	8232.779	2.85	35.24	61.91	
	4	-3984.643	48	8283.806	47.65	7.56	1.28	43.51
t_{PTEF}/t_E	1	-4023.865	14	8139.466	100.00			
	2	-4005.411	26	8181.187	20.68	79.32		
	3	-3992.756	38	8234.508	2.71	29.24	68.05	
	4	-3985.912	48	8286.345	1.28	47.36	9.84	41.51

Table 4.22 presents the model estimates for resistance, compliance, Ree, and Xee from the LCMM models using splines. The table includes effect sizes with corresponding 95% CI and p-values across two latent classes.

	FRC		LCI		Respiratory rate		Tidal volume		t_{PTEF}/t_E	
	Effect size (95% CI)	**p-value**	**Effect size (95% CI)**	**p-value**	**Effect size (95% CI)**	**p-value**	**Effect size (95% CI)**	**p-value**	**Effect size (95% CI)**	**p-value**
Intercept Class 1 (not estimated)	0		0		0		0		0	
Intercept Class 2	-1.394 (-2.258, -0.53)	0.002	-1.802 (-3.418, -0.185)	0.029	-1.296 (-1.59, -1.002)	0.000	-0.988 (-1.586, -0.39)	0.001	-0.867 (-1.258, -0.476)	0.000
bs 1 Class 1	-2.257 (-4.34, -0.174)	0.034	-0.01 (-0.434, 0.413)	0.962	-1.152 (-1.713, -0.59)	0.000	-0.207 (-1.115, 0.701)	0.655	-0.264 (-1.254, 0.726)	0.601
bs 1 Class 2	-0.061 (-0.589, 0.468)	0.822	8.311 (4.653, 11.968)	0.000	0.55 (-0.049, 1.15)	0.072	0.256 (-0.265, 0.777)	0.335	0.252 (-0.24, 0.744)	0.315
bs 2 Class 1	0.774 (-1.259, 2.807)	0.455	-0.107 (-0.426, 0.213)	0.513	-0.559 (-1.078, -0.04)	0.035	-1.358 (-2.344, -0.371)	0.007	-1.384 (-2.423, -0.346)	0.009
bs 2 Class 2	0.111 (-0.413, 0.636)	0.677	-0.454 (-3.027, 2.119)	0.730	0.84 (0.275, 1.406)	0.004	0.324 (-0.553, 1.202)	0.469	0.255 (-0.518, 1.028)	0.518
bs 3 Class 1	-2.872 (-4.741, -1.004)	0.003	0.084 (-0.138, 0.306)	0.458	-1.329 (-1.816, -0.841)	0.000	-1.013 (-2.156, 0.129)	0.082	-1.144 (-2.359, 0.072)	0.065
bs 3 Class 2	0.542 (0.041, 1.042)	0.034	3.708 (1.693, 5.724)	0.000	1.697 (1.259, 2.135)	0.000	0.752 (0.209, 1.296)	0.007	0.711 (0.201, 1.221)	0.006
Child characteristics										
Sex (male) Class 1	-0.019 (-0.321, 0.283)	0.903					0.095 (-0.056, 0.246)	0.216		
Sex (male) Class 2	0.092 (-0.018, 0.202)	0.103					0.113 (0.011, 0.215)	0.030		
Preterm (<37weeks) Class 1	-0.202 (-0.721, 0.317)	0.446			0.1 (-0.047, 0.246)	0.181	-0.131 (-0.366, 0.104)	0.274	-0.122 (-0.378, 0.135)	0.352
Preterm (<37weeks) Class 2	0.055 (-0.106, 0.216)	0.505			0.086 (-0.066, 0.239)	0.268	-0.053 (-0.191, 0.084)	0.447	-0.057 (-0.189, 0.076)	0.401
HIV exposed uninfected versus HIV unexposed Class 1			-0.042 (-0.138, 0.054)	0.392			-0.033 (-0.212, 0.146)	0.721		
HIV exposed uninfected versus HIV unexposed Class 2			-0.507 (-1.543, 0.53)	0.338			-0.022 (-0.153, 0.109)	0.741		
Birthweight-z-score Class 1	0.009 (-0.142, 0.159)	0.912			-0.006 (-0.056, 0.044)	0.823	0.03 (-0.047, 0.108)	0.442	0.023 (-0.06, 0.105)	0.594
Birthweight-z-score Class 2	0.005 (-0.048, 0.059)	0.846			-0.031 (-0.082, 0.019)	0.225	0.021 (-0.028, 0.071)	0.394	0.014 (-0.033, 0.061)	0.553
Wheezing phenotype										
Early transient (vs Never) Class 1			0.105 (0.009, 0.201)	0.032	0.012 (-0.131, 0.155)	0.871			-0.114 (-0.318, 0.09)	0.274
Early transient (vs Never) Class 2			0.122 (-0.876, 1.12)	0.810	-0.017 (-0.158, 0.124)	0.812			-0.019 (-0.144, 0.106)	0.765
Late onset (vs Never) Class 1			-0.028 (-0.15, 0.094)	0.650	-0.104 (-0.274, 0.067)	0.233			-0.077 (-0.345, 0.191)	0.573
Late onset (vs Never) Class 2			-0.653 (-1.978, 0.672)	0.334	-0.028 (-0.223, 0.166)	0.776			0.035 (-0.126, 0.197)	0.667
Recurrent (vs Never) Class 1			0.088 (-0.018, 0.195)	0.103	-0.035 (-0.195, 0.126)	0.674			-0.08 (-0.314, 0.153)	0.500
Recurrent (vs Never) Class 2			0.218 (-0.562, 0.999)	0.583	-0.105 (-0.28, 0.069)	0.236			0.037 (-0.105, 0.179)	0.606
LRTI										
LRTI ever Class 1	-0.206 (-0.527, 0.115)	0.208			0.057 (-0.062, 0.176)	0.347				
LRTI ever Class 2	-0.003 (-0.117, 0.11)	0.954			0.061 (-0.064, 0.186)	0.339				
Maternal characteristics										
Postnatal IPV Class 1	0.203 (-0.139, 0.544)	0.245							0.118 (-0.077, 0.314)	0.234
Postnatal IPV Class 2	0.076 (-0.048, 0.2)	0.228							0.036 (-0.076, 0.148)	0.531
Maternal asthma/allergy Class 1					-0.004 (-0.179, 0.171)	0.968				
Maternal asthma/allergy Class 2					0.149 (-0.06, 0.359)	0.162				
Postnatal smoking Class 1	-0.101 (-0.433, 0.231)	0.552			-0.045 (-0.158, 0.068)	0.438			-0.023 (-0.193, 0.147)	0.792
Postnatal smoking Class 2	-0.032 (-0.148, 0.083)	0.586			-0.078 (-0.196, 0.039)	0.190			-0.051 (-0.158, 0.056)	0.350
Socio-economic status (SES)										
Asset ownership										
Low-Medium (vs Low) Class 1					-0.035 (-0.186, 0.116)	0.650				
Low-Medium (vs Low) Class 2					-0.060 (-0.209, 0.090)	0.434				
Medium-High (vs Low) Class 1					-0.050 (-0.207, 0.106)	0.529				
Medium-High (vs Low) Class 2					-0.101 (-0.250, 0.048)	0.186				
High (vs Low) Class 1					-0.123 (-0.279, 0.033)	0.122				
High (vs Low) Class 2					-0.100 (-0.264, 0.064)	0.234				
Maternal eduction										
Secondary (vs Primary) Class 1							-0.149 (-0.444, 0.145)	0.320		
Secondary (vs Primary) Class 2							-0.045 (-0.246, 0.156)	0.664		
Completed secondary (vs Primary) Class 1							-0.128 (-0.446, 0.190)	0.429		
Completed secondary (vs Primary) Class 2							0.035 (-0.178, 0.248)	0.746		
Any tertiary (vs Primary) Class 1							-0.224 (-0.656, 0.207)	0.308		
Any tertiary (vs Primary) Class 2							-0.006 (-0.318, 0.305)	0.970		

Table 4.24: Model estimates for FRC, LCI, respiratory rate, tidal volume, and t_{PTEF}/t_E from LCMM.

Previously strong associations weaken

Previous strong associations observed in earlier analysis in this chapter, when modelling single trajectories, could weaken when using LCMM for several reasons. In single trajectory models, the cohort is treated as homogeneous, which leads to an averaged effect of covariates, potentially masking subgroup differences. When using LCMM, the population is divided into latent classes, and each class may have different patterns of association. A covariate that appeared strongly associated with the single trajectory model might not be significant in some or all latent classes, indicating that the effect was driven by certain subgroups, not the entire population. Additionally, transforming outcomes into z-scores can further weaken these original associations. Standardisation emphasises relative differences rather than absolute changes, which can dampen the strength of associations, especially if the predictors had a more direct influence on the raw values of the outcome. Transforming to z-scores also compresses the variability of the data slightly, by reducing the spread to a common distribution. This compression can weaken the apparent relationship between predictors and the standardised outcome. Z-scores put more emphasis on relative differences from the mean. In cases where predictors drive large changes in the outcome (e.g., strong positive or negative effects in raw values), converting to a z-score can make these large effects less apparent, reducing the observed strength of associations.

Another important consideration is that the associations are being compared between linear and spline models, which differ in complexity. Spline models provide a high degree of flexibility, where the model closely follows the noise or minor fluctuations in the data, possibly reducing the overall strength of the associations. This contrasts with linear models, which provide a more constrained fit, potentially leading to stronger (but less nuanced) associations because the relationships are forced into a simpler form. While splines can capture subtle, non-linear patterns, they may also smooth out extreme changes that would otherwise be captured as strong associations in linear models. This smoothing effect can lead to weaker overall associations, as the model distributes the impact of predictors across multiple knots or segments.

Class-specific predictions

This section explores the development of the latent class structure as the number of classes increases from one to four using class-specific predictions. As the number of latent classes increases, different patterns of lung function emerge, suggesting that children may follow distinct lung function trajectories based on underlying factors. Figures 4.26 and 4.27 present the z-score trajectories across different latent class solutions with age. These trajectories compare groups of children based on several factors, including a history of LRTI, sex differences, HIV exposure, and preterm versus full-term births. The figures show how covariates, like LRTI status, impact lung function outcomes within each class. The presence or absence of LRTI can shift the trajectory upwards or downwards within each class.

With regards to oscillometry outcomes, the findings from the two-class solutions in Table 4.23

showed children in Class 2 had higher baseline values for Ree (0.507, 95% CI [-0.004, 1.018], p = 0.052) and Xee (0.433, 95% CI [0.013, 0.852], p = 0.043).

The basis spline estimates reveal complex relationships between the predictors and the outcomes of resistance, compliance, Ree, and Xee across two latent classes. In Class 1, positive effects are observed for the first spline basis (bs 1), indicating that higher values correspond to increased resistance (effect size = 3.182) and compliance (effect size = 1.156), with p-values below 0.001. Conversely, Class 2 shows negative effects for bs 1, suggesting a decrease in resistance (effect size = -3.299) and compliance (effect size = -4.752) as predictor values rise, also statistically significant (p < 0.001). The second basis spline (bs 2) demonstrates contrasting effects across classes, with Class 1 experiencing a negative association for resistance and compliance, while Class 2 shows positive associations for both outcomes. The estimates for the third basis spline (bs 3) exhibit weaker associations, with significant positive effects for resistance in Class 1 (effect size = 0.362) but no significant effects for Class 2.

With regards to MBW and tidal breathing outcomes, the findings from the two-class solutions in Table 4.24 showed that children in Class 2 had significantly lower intercepts for FRC (-1.394, 95% CI [-2.258, -0.53], p = 0.002), LCI (-1.802, 95% CI [-3.418, -0.185], p = 0.029), respiratory rate (-1.296, 95% CI [-1.59, -1.002], p < 0.001), tidal volume (-0.988, 95% CI [-1.586, -0.39], p < 0.001), and t_{PTEF}/t_E (-0.867, 95% CI [-1.258, -0.476], p < 0.001). This suggests that children in Class 2 had consistently lower baseline values across multiple lung function metrics compared to Class 1.

FRC in Class 1 decreases over time, as shown by bs 1 Class 1 (-2.257, 95% CI [-4.34, -0.174], p = 0.034) and bs 3 Class 1 (-2.872, 95% CI [-4.741, -1.004], p = 0.003). LCI in Class 2 shows a strong increase over time with a significant effect for bs 1 Class 2 (8.311, 95% CI [4.653, 11.968], p < 0.001) and bs 3 Class 2 (3.708, 95% CI [1.693, 5.724], p < 0.001). This indicates worsening lung clearance function over time in Class 2. Respiratory rate shows a decline in Class 1 over time (bs 1 Class 1, -1.152, 95% CI [-1.713, -0.59], p < 0.001) and bs 3 Class 1 (-1.329, 95% CI [-1.816, -0.841], p < 0.001). Tidal volume in Class 2 increases over time, with bs 3 Class 2 showing a positive effect (0.752, 95% CI [0.209, 1.296], p = 0.007), while there is no change in Class 1. The spline terms indicate sharper changes at the beginning and end of the age domain, while some effects are less pronounced in the middle. This could reflect critical periods in lung development when children are particularly vulnerable to external insults (e.g., infections, and environmental exposures).

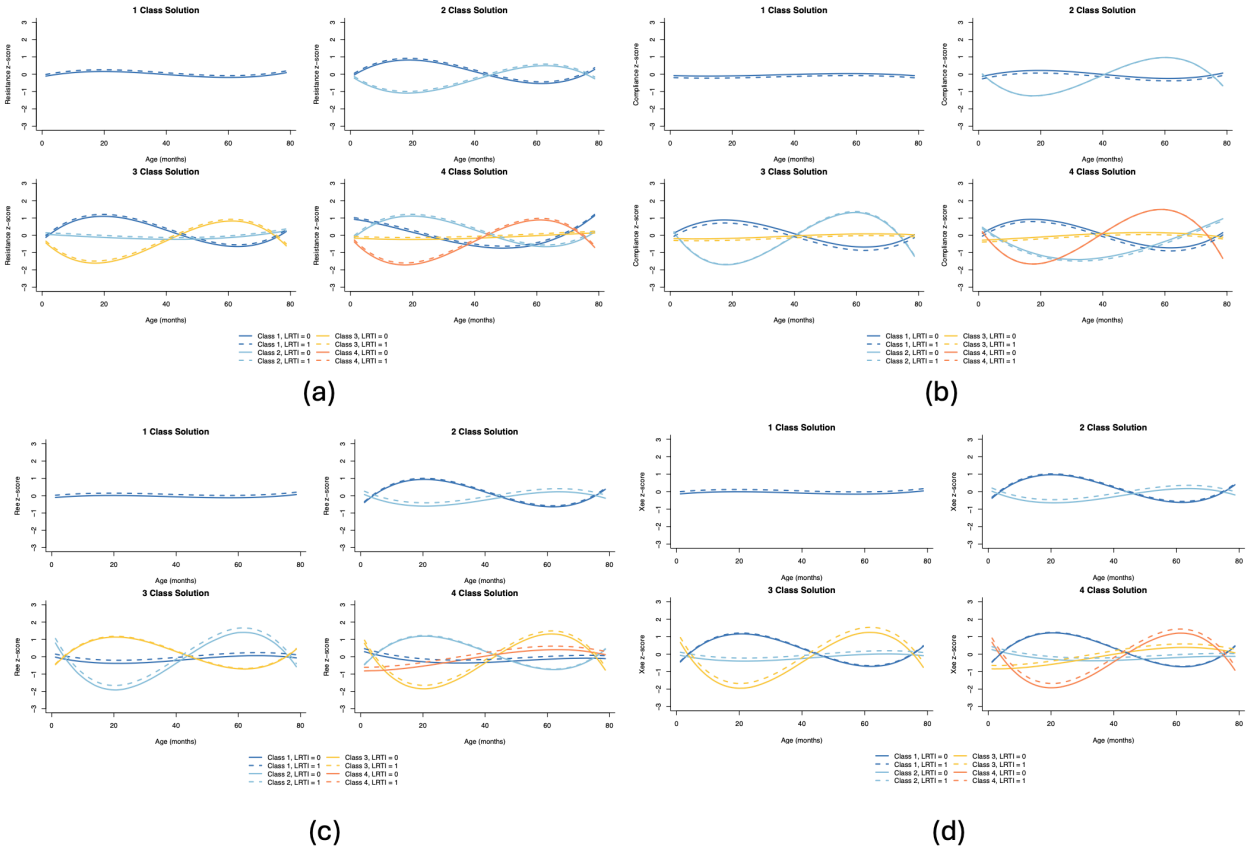


Figure 4.26: This figure illustrates the heterogeneous structures within airwave oscillometry z-scores using LCMM with spline models.

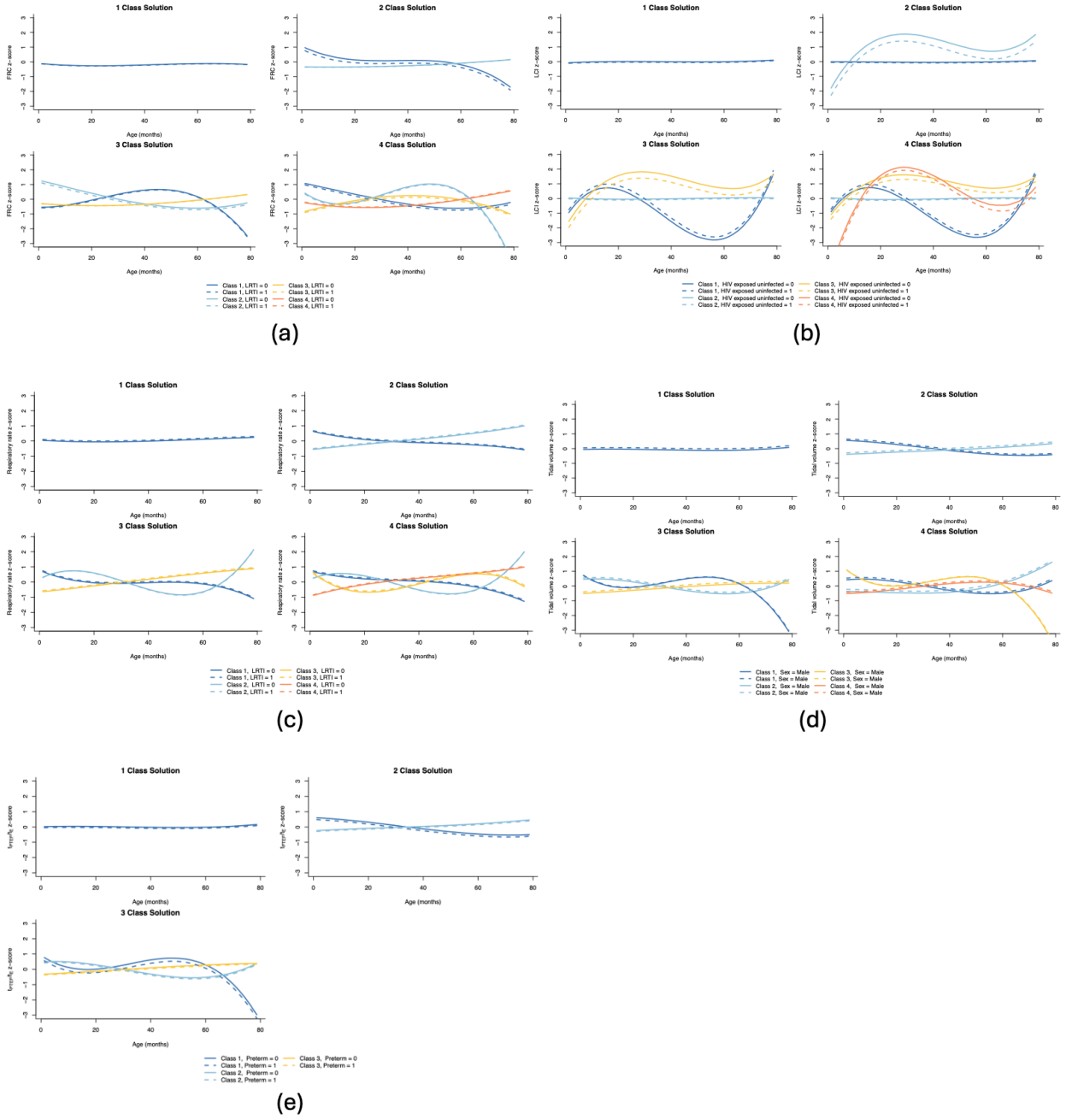


Figure 4.27: This figure illustrates the heterogeneous structures within MBW and tidal breathing resistance z-scores using LCMM with spline models.

Class separation

Low class separation (i.e. high overlap between classes) across all outcomes is a factor which should be considered under model performance and should be minimised where possible. Class separation can be quantified by assessing the overlap in the intercepts and slopes of the latent classes' growth trajectories, or by evaluating the extent of overlap in the observed repeated measures, [167]. Trajectories are defined by parameters like the intercept (starting level) and slope (rate of change over time). Clear distinctions in these parameters are helpful for validating that each class represents a distinct pattern. Ideally, each class should represent a group that behaves differently from the others in a meaningful way, whether this difference is based on growth rate, starting point, or pattern of change over time.

The oscillometry results from Table 4.22, indicate intercepts for both resistance and compliance in Class 2 show minimal class distinction from those in Class 1. For resistance, Class 1 and Class 2 show strong slope differentiation (bs 1 and bs 2). For compliance, differences exist in bs 1 and bs 2 between classes. For Ree, the Class 2 intercept indicates near overlap with Class 1, while differences in slopes (bs 1) and (bs 2) suggest some separation between the classes. For Xee, Class 2's intercept is different to that of Class 1, with opposing significant slopes between classes. The bs 3 term across all four outcomes suggests a lack of separation at the top end of the age range. This pattern highlights that class distinctions diminish at older ages, potentially indicating convergence in outcomes over time. The oscillating pattern observed in the plotted trajectories of the latent classes, in Figure 4.26, further underscores this, as oscillations within the slopes can contribute to reduced interpretability and overlapping intervals between classes. This oscillatory behaviour suggests that distinct separation is not sustained over time.

The MBW and tidal breathing results from Table 4.24, indicate for FRC, Class 2 has a lower intercept, indicating some separation, but partial overlap in slopes was observed. In LCI, Class 2 showed separation in intercept and some slopes, indicating divergence despite some overlap. For respiratory rate, Class 2's lower intercept suggests separation, but mixed slope significance indicates overlap. Tidal Volume shows intercept separation but mixed slope separation, suggesting potential overlap. Lastly, t_{PTEF}/t_E shows a lower intercept for Class 2 with separation at the top end of the age range.

As the number of classes increases, a greater overlap between latent classes naturally occurs. This is evident in Figures 4.26 and 4.27. More classes dilute the separation in parameters (intercepts and slopes) since the range of observed values for each parameter must be distributed among a greater number of classes. Clinically, this increased overlap may complicate the interpretation of class-specific trajectories, as the subtle differences between classes might not represent clinically meaningful subgroups but noise due to the over-division of the sample.

4.4 Discussion

South Africa faces a significant challenge in addressing respiratory health due to severe inequality and resource-limited settings. Like other LMIC countries with substantial poverty, South Africa bears a disproportionate burden of respiratory diseases, which primarily affect its poorest communities.

This chapter completed a detailed analysis of the longitudinal trajectories of lung function in a South African birth cohort of almost 1000 children. The trajectories for multiple indices, including oscillometry, tidal breathing, and MBW, were described, along with several factors associated with those trajectories. This study extends previous findings in the same cohort at ages 1 and 2 years and adds to the understanding of factors linked with poorer lung function in LMIC.

The impact of height on lung function development is well-recognised in paediatric respiratory research, given its strong association with many lung function outcomes. The early parts of this chapter's analysis confirm that height is a significant factor influencing most lung function. The use of a multiplicative model to adjust for height allows for a more accurate assessment of the underlying respiratory function independent of body size variations. The consistency of model coefficients between height-adjusted and unadjusted models underscores the robustness of these modelling approaches. Whether height is included directly as a predictor or adjusted out from the lung function outcome, the model estimates remain largely stable, supporting the validity of either method in paediatric lung function analysis.

Comprehensive changes in lung development during infancy and the preschool years were thoroughly described, along with the impact of multiple exposures on lung growth. Known factors such as maternal smoking and asthma were found to negatively affect lung development, along with other risks like HIV exposure and poorer households. Early-life LRTI impaired lung growth and development, with reduced respiratory system compliance, and increased resistance and respiratory rate in those with a previous LRTI. Recurrent LRTI was associated with additional impairment for each LRTI. RSV, the leading global cause of childhood LRTI, had the most significant impact, particularly in children hospitalised for RSV-LRTI, who showed lower expiratory flow ratios than those with non-RSV-LRTI. In contrast, RV-LRTI had a similar impact to non-RSV-LRTI.

It was previously reported in the DCHS that early-life LRTI impacted lung function to two years, independent of baseline lung function, [65, 24]. This analysis extends previous findings by demonstrating that the impact of early-life LRTI, even an ambulatory event, persists and increases through to age five years. These findings align with studies from HICs showing that childhood LRTI is linked to reduced spirometric lung function in later life, [17, 66]. This analysis has shown that, in South Africa, the impact of RSV on lung development is substantial, surpassing the effect of RV-LRTI, which is the virus most strongly linked to childhood obstructive airway disease in HICs. RSV-LRTI requiring hospitalisation was found to have a greater impact on lung development compared to ambulatory RSV-LRTI. However, this pattern did not extend to all-cause or RV-LRTI. Additionally,

children with RSV-LRTI exhibited a lower t_{PTEF}/t_E trajectory compared to those with non-RSV LRTI, consistent with the development of airway obstruction following early-life RSV-LRTI, [19, 25]. This study highlights the need for RSV prevention, especially in high-burden settings such as South Africa, and the need to prioritise careful follow-up of children hospitalised with RSV-LRTI.

Novel evidence was provided through the analysis which showed that in addition to aetiology and severity, it is the recurrence and timing of early-life LRTI that worsen its impact on lung development. Recurrent events were associated with additional adverse impacts on lung volumes, but also an increase in LCI, suggesting early peripheral airway damage. If LRTI occurred in the first year of life, the sustained effect was larger than if occurring later, suggesting that lung growth effects are strongest with early harm.

The progression of clinical findings is illustrated in this chapter through the integration of wheezing phenotypes identified in the previous chapter. Where the wheezing phenotypes were used as predictors of lung function trajectories under the GAMLSS approach in combination with an ITS approach. The lung function measurements exhibited varying associations with wheeze phenotypes, offering insights into the underlying mechanics. Late-onset wheeze was associated with an increase in respiratory resistance over time, consistent with developing reduced airway calibre. Recurrent wheeze through childhood was associated with higher resistance and lower compliance, expiratory flow ratio and respiratory rate throughout pre-school years, consistent with persistent airflow limitation.

HIV exposure increases the risk for LRTI in infancy and alters lung function at birth, [64]. Previous research within the DCHS cohort has demonstrated that HEU children whose mothers had more severe HIV disease showed impairments in LCI and resistance by the age of two, [162]. The current study extends this finding by showing that in children HEU, lung function development is impaired from birth to school age, with increases in LCI showing a sustained effect of HIV exposure throughout childhood in children whose mothers had a detectable HIV viral load during pregnancy.

Preterm birth, particularly before 32 weeks of gestation, is known to have long-term effects on respiratory health, [168]. In addition, moderate to late preterm birth (32 to <37 weeks) might affect long-term respiratory health, [169]. The analysis in this chapter reinforces existing research, showing that children born preterm in the DCHS cohort experienced decreased lung function over time. This association between preterm birth and lung function is particularly important given the global burden of preterm birth and the effect this could have on long-term morbidity and mortality.

Postnatal maternal smoking was associated with altered development of lung function consistent with airway obstruction. Prenatal maternal psychological adversity, including maternal IPV, has been associated with increased risk of childhood wheeze and early-life LRTI,[170] and, as previously described in this cohort, impaired lung function in infancy, [135]. Additionally, indirect effects of poorer maternal dietary and lifestyle habits during pregnancy could contribute to impaired child lung growth, [135, 171]. The analysis completed in this chapter shows the effect of postnatal

maternal IPV on lung function development, including impaired expiratory flow ratio t_{PTEF}/t_E, up to school age, which is consistent with the previously observed increased risk of wheeze within the DCHS cohort, [25].

This analysis has also shown that multiple exposures appear additive, with children with co-existing risk factors having the poorest lung development. This finding is important in risk stratification and prioritising preventive interventions and management planning.

Regarding the novelty and significance of the clinical findings, key insights include the influence of both the timing and severity of LRTIs. Notably, even mild LRTIs occurring within the first year of life have a lasting impact, persisting at least through school age. This is important as the lung function when children first enter school is strongly predictive of the physiological peak reached in early adulthood. Additionally, the comprehensive description of lung development from birth through the preschool years fills an important gap in current trajectory data and literature, enhancing our understanding of early lung growth and development.

Limitations of this study include that adequately adjusting for the timing of multiple exposures is challenging in an observational cohort. In addition, multiple exposures and comparisons might lead to spurious findings; however, an adjustment was made for multiple comparisons and the associations remained.

From a statistical viewpoint, the study showed how GAMLSS can be used to model the development of lung function. The flexibility of GAMLSS was demonstrated through The LMS method which was applied to standardise anthropometric measures. Additionally, the GAMLSS framework, combined with an ITS approach, was used to explore the early-life development of lung function. This allowed for modelling complex, growth patterns and investigating how early-life factors impact lung function trajectories. However, one concern while using GAMLSS is that it is an extremely flexible framework which can lead to complex models. Care needs to be taken to reduce the risk of overfitting. Although GAMLSS offers more than 40 distribution options, the distributions used here (BCCG, BCPE, and normal) are well-established, which helps minimise the uncertainty in model selection, [172].

Using LCMM, it was observed how the heterogeneous structures within the data can be explained by exploring different latent class structures. In general, two trajectories were identified as the best explanation for the latent class structure. These latent classes differ in size, with the majority of the cohort following a stable, near-normal trajectory, while a smaller class represents children who deviate from this pattern.

LCMMs are of practical importance since by identifying distinct groups, a better understanding may arise of differences in development between groups and/or risk factors (e.g. by linking to covariates such as smoke exposure). This thesis studied both univariate (i.e. considering one outcome) and multivariate (i.e. considering multiple measures simultaneously) outcomes. Univariate LCMMs allowed for a detailed examination of specific outcomes and were useful in identifying unique class trajectories for individual measures, highlighting particular outcomes that may be sensitive indi-

cators of atypical development. By extending to multivariate models, it was possible to examine shared latent processes that influence multiple dimensions of lung function. This is beneficial in contexts where outcomes are expected to be interdependent due to underlying physiological or environmental factors. The multivariate LCMM approach incorporates correlations among multiple outcomes, which enhances the stability and interpretability of the latent classes.

How risk factors are incorporated into an analysis can vary. The LCMM approach directly incorporates multiple risk factors into the latent class formation process through the mixture component. This method acknowledges that latent classes arise from underlying biological and environmental differences, enabling factors like sex or LRTI to exert variable influences across classes. In contrast, another option is the cluster-then-analyse approach, as used by Granell et al. (2024) [80], which involves first identifying latent classes based solely on lung function trajectory data and then using statistical models, such as multinomial logistic regression, to associate early-life risk factors with these predefined classes. This method separates class identification from the analysis of risk factors. In their paper, Granell et al. (2024) used the same cluster-then-analyse approach to identify latent class specific risk factors. After first identifying latent classes of trajectories, multinomial logistic regression models were then used to identify early-life risk factors associated with each lung function trajectory. The cluster-then-analyse method emphasises ease of interpretation at the cost of capturing nuanced covariate-class interactions.

The analysis around class separation reinforced the importance of careful interpretation as it highlights issues with low class separation that complicate decisions about the number of latent classes. Specifically, the separation between classes was low enough to question whether increasing or decreasing the number of classes would yield more meaningful distinctions. This leads to two possible interpretations when low separation is observed among extracted classes: either fewer classes would be more interpretable, capturing only the most substantive distinctions, or additional classes might be needed, despite further reducing separation, to capture subtle structures in the data fully. Therefore, the findings correspond with the findings by van der Nest (2022) that low separation could either indicate a need for fewer classes to maintain substantive differentiation or point to the possibility of missing classes, though adding more may not enhance clinical relevance.

Convergence of LCMM models was not a barrier in the univariate analysis within the DCHS data, however, model convergence needed to be addressed in some scenarios for multivariate outcomes. It was found that in some cases the estimation process reached the maximum number of iterations (set at 100) without achieving an optimal solution. More complex models were more susceptible to non-convergence. In instances of non-convergence, a grid search function proved helpful when it was believed that the optimiser was getting stuck at a local minimum. Proust (2017) warns that estimates on the boundaries of the parameter space could be the reason for non-convergence.

Due to the granularity, timing, and completeness of early-life lung function measurements in the DCHS, it can be challenging to find comparable studies for each of the nine outcomes. However, a 3-year observational study in the UK of repeated LCI measurements in patients under 5 years

provides a possible comparison, [78]. The latent class growth analysis using LCMM and a sample of 81 children revealed a stable, near-normal trajectory, which contained the majority of children. This stable, near-normal class strongly agrees with the structure identified in the DCHS. Further, smaller classes of increasing trajectories were observed by Horsley et al. (2022). A smaller, increasing LCI class was also observed in the DCHS, however, the baseline values of the increasing classes relative to the stable, near-normal class differ between the DCHS and Horsley et al. (2022). In the DCHS the increasing classes of LCI have a lower baseline value compared to stable, near-normal class. Whereas, Horsley et al. (2022) reported higher baseline values for the increasing classes of LCI. It should be noted that Horsley et al. (2022) implemented LCMM using a quadratic model structure. Given the smaller sample size used by Horsley et al. (2022), the analyses should be regarded as exploratory.

Further approaches are available when analysing latent growth profiles, such as group-based trajectory modelling (GBTM) - a version of growth mixture models - which also allows researchers to identify clusters of individuals who follow similar trajectories for a single outcome of interest. GBTM requires two main inputs, 1) a function defining the shape of the trajectory (for example, including age as a non-linear predictor using a polynomial function) and 2) the probability of trajectory class membership. GBTM has further been extended to deal with dual group-based trajectory modelling, [173]. The dual GBT model jointly estimates the trajectories of two related longitudinal outcomes, where the two outcomes can evolve in a coeval manner or they can evolve over different time periods. However, this analysis does not scale well to more than two trajectories because the number of conditional probability tables linking the different outcomes quickly becomes unwieldy, [174]. This problem has been addressed by Nagin & Jones (2018)[174], through a multi-trajectory model that accompanies analyses relating trajectories for more than two outcomes.

Chapter 5

Conclusion and unifying wheezing and lung function phenotypes insights through correspondence analysis

5.1 Introduction

5.1.1 Concluding summary

In this thesis, the analysis focused on the South African DCHS birth cohort, a unique population from an LMIC. The study brought together two extensive segments exploring early childhood development, namely wheezing phenotypes and lung function trajectories.

This thesis addressed the comprehensive process of examining the underlying heterogeneity in wheezing and lung function measurements. This included data transformation and standardisation, selection of dissimilarity functions, comparison of clustering methods, determination of the optimal number of classes, stability analysis, assessment of the impact of different input variables across approaches, and clinical interpretation. Through these steps, the aims of the thesis were achieved by successfully identifying clinically meaningful subgroups and uncovering factors associated with distinct wheezing and lung function trajectories.

Chapter 1 introduced the setting, study, design, and participants of the DCHS. In chapter 2, a literature review discussed different major methodological approaches to longitudinal clustering problems, and their evolvement over time, with reference to different cohorts. During the literature review of wheezing phenotypes, it was found that the PAM approach tailored for longitudinal data was identified as having distinct strengths, as it integrates both statistical and clinical patterns of thought through multidimensional indicators. In recent years, this approach has increasingly been applied in studies as an alternative to LCA, which has traditionally been considered the gold standard. The PAM approach was notably employed in the UK ALSPAC cohort to derive

wheezing phenotypes, producing the benefit of a more direct comparison with the DCHS. The review on longitudinal trajectories of lung function showed that the LCMM method, used in several high-income cohorts, has enabled the classification of multiple lung function trajectories and their associations with specific risk factors. Given LCMM's successful application in HICs for identifying distinct lung function patterns, the method was identified as a viable approach within the DCHS. The literature review highlighted the importance of understanding the closely connected relationship between wheezing phenotypes and lung function, as consistently highlighted across studies.

The research in Chapter 3 presented the first comprehensive study of wheezing phenotypes, risk factors, and their impact on lung function in a birth cohort from an LMIC, following children from birth to age 5. Key findings include the identification of four distinct wheezing phenotypes using the PAM clustering approach tailored for longitudinal data, offering greater clinical interpretability than LCA. The risk factor findings from Chapter 3 highlight the importance of region-specific factors in shaping wheezing trajectories. To the best of current knowledge, this is the first study in an LMIC context to have derived wheeze phenotypes using an unsupervised method. In contrast to other birth cohort studies, the large number of data collection time points (14) in the first five years of life has enabled the focus to intensively be on early childhood. A major strength of the study is the derivation of internally homogeneous phenotypes, and the comparison with ALSPAC has enabled a direct comparison of the internal structure of the phenotypes. No such comparison exists in previous LCA studies despite phenotypes sharing similar nomenclature. While selection bias could occur at baseline, it is not considered a major methodological limitation in this study. Although some selection bias could exist due to excluded participants having higher SES, a factor linked to higher wheezing prevalence but lower maternal smoking rates compared to included participants, the impact is likely minimal. The high retention rate, along with the inclusion of most participants who completed comprehensive data collection across 14 time points, reduces the likelihood of substantial bias. Additionally, the primary reason for loss to follow-up was relocation, which further minimises concerns about selection bias. Given the exploratory nature of cluster analysis, there are no clear guidelines on the sample size requirements or the relationship between the number of clusters and the number of clustering variables used. A simulation study found that increasing the sample size from 10 to 30 times the number of clustering variables substantially improves the quality of the clustering solution. While the benefits decrease with further increases, improvements continue up to a sample size of 100 times the number of variables, [175]. Therefore, given 6 indicator variables, a sample size of 950 children is acceptable for clustering. The stability of the cluster solution was further tested by simulating changes in sample size, during which the optimal solution was found to be stable.

The research in Chapter 4 extended and confirmed previous findings within the DCHS, highlighting factors linked to impaired lung development. The influence of height on lung function was confirmed in Chapter 4. The findings highlight the statistical consistency and robustness of the height-adjusted approach, affirming that the height adjustment can be applied without altering model interpretability or accuracy. It is important to clarify further the rationale behind seeking to

account for the impact of body size. There are several lung function measurements which depend on both body size and age and are standard considerations in longitudinal lung growth trajectories. In this thesis, height (as a proxy for body size) adjusted variables were used in the model as previously described and previously used in lung function trajectory analyses, [70]. Using this multiplicative model approach, the specific influence on lung function was isolated, distinguishing it from variations in height, an established predictor of certain lung function measures. It is acknowledged that height is only one component of body size and that, for example, body mass has an important impact on lung development. Collinearity of weight-for-height or BMI is a challenge in this model which includes multiple exposures. The two-class solutions using GAMLSS and LCMM were statistically optimal, balancing model fit and interpretability while minimising class overlap, effectively capturing complex lung growth patterns and identifying latent classes. The findings highlight unique regional risks, illustrating the additive impact of multiple exposures on lung development and informing targeted preventive strategies.

5.1.2 Bringing the different chapters together

Up to this point, the discussion has approached causal associations with caution, yet associations still exist. In the preceding chapters, the relationship between wheezing phenotypes and lung function was modelled using different setups. Chapter 3 aimed to identify how lung function can predict the occurrence and severity of wheeze phenotypes. Conversely, Chapter 4 investigated how the presence of different wheeze phenotypes influences lung function outcomes, providing insights into how wheezing patterns may impact respiratory health over time.

In this section, the analysis from prior chapters is brought together by using multiple correspondence analysis (MCA) to analyse the association between wheezing phenotypes and lung function phenotypes. A key attribute of correspondence analysis lies in its ability to visually represent the relationships between rows and columns through biplots, facilitating the identification of underlying structural patterns. This is done by approximating the given data in a lower dimension which allows the associations between wheezing phenotypes, between lung function classes, and between wheezing phenotypes and lung function classes to be investigated. To maintain consistency throughout this thesis, the analysis was conducted separately for the three different sets of outcomes: airwave oscillometry, tidal breathing, and MBW measures.

5.1.3 Multiple correspondence analysis (MCA)

Simple correspondence analysis (SCA) is a technique for exploring associations within a two-way contingency table by examining the relationships between row and column categories, with a focus on their relative values. The solution to SCA, as described by Greenacre (1984), is effectively represented through the singular value decomposition of a transformed matrix, [176]. This decomposition yields principal coordinates that allow for the low-dimensional representation of categorical relationships, making it particularly useful for visualising complex associations. MCA is essentially the application of the SCA algorithm to multivariate categorical data coded in the form of an indicator

matrix or a Burt matrix, [176, 177]. Therefore two computational methods exist, computations based on the indicator matrix or computations based on the Burt matrix.

The most classical approach to MCA, which is used in this chapter, is to apply simple correspondence analysis to the indicator matrix $\mathbf{Z}$. Initially, a dummy (or indicator) variable is defined for each of the Q categorical variables that comprise the table. The q^{th} variable has the J_q categories and $J = \sum_{q=1}^{Q} J_q$ is the total number of categories over all variables. Given that there are n individuals in the study, define the matrix $\mathbf{Z} : (n \times J)$ with ij^{th} element equal to

$$Z_{ij} = \begin{cases} 1, & \text{if the } i^{th} \text{ individual belongs to the } j^{th} \text{ category} \\ 0, & \text{otherwise,} \end{cases} \tag{5.1}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, J$. It is assumed that an individual belongs to at least one category of each of the Q categorical variables. Each row sums to Q and the sum of all entries is nQ. The matrix $\mathbf{Z}$ can be written as

$$\mathbf{Z} = (\mathbf{Z}_1 \ldots \mathbf{Z}_Q)\,, \tag{5.2}$$

where $\mathbf{Z}_q$ is the $(n \times J_q)$ matrix corresponding to the q^{th} categorical variable. MCA is then a simple correspondence analysis applied to the partitioned matrix $\mathbf{Z} = (\mathbf{Z}_1 \ldots \mathbf{Z}_Q)$, and the resulting decomposition can reveal structures that provide insight into the complex associations between multiple categorical variables.

5.1.4 Discussion

The analysis reduces the high-dimensional categorical data into a few principal dimensions. Each lung function class and wheezing phenotype is positioned on a biplot based on their contributions to the first two dimensions. The MCA biplot in Figure 5.1 displays two dimensions, dimension 1 and dimension 2, which are the principal components derived from the MCA. These dimensions represent the directions of maximum variance in the data. Across all analyses, dimension 1 captures a significant amount of variance, driven by specific lung function outcomes which may relate to structural aspects of the lungs, while dimension 2 generally explains phenotypic variation (wheezing patterns). The MCA results show largely 2 independent components.

It is important to note that the plotted points displayed in the MCA biplot in Figure 5.1 represent estimated positions based on model outputs from chapters 3 and 4, rather than raw data. Consequently, any interpretation of variance or associations between lung function classes and wheezing phenotypes should take into consideration the inherent estimation uncertainty. While the first two dimensions capture the major sources of variation, the variability of lung function outcomes should be understood within the context of model-driven approximations rather than direct empirical measures.

Additionally, since the MCA model is estimated from a finite sample, there is inherent uncertainty in the positioning of categories. In other words, if this analysis was repeated on a different sample from the same cohort, the exact positions of the points might shift slightly. Finally, the variance captured by dimension 1 and dimension 2 reflects the proportion of total inertia (i.e., variance) explained by these dimensions, but this does not mean it directly measures the true biological variance of lung function. Instead, it shows how much of the variation in categorical responses (i.e., lung function classes and wheezing phenotypes) is represented by these two principal dimensions.

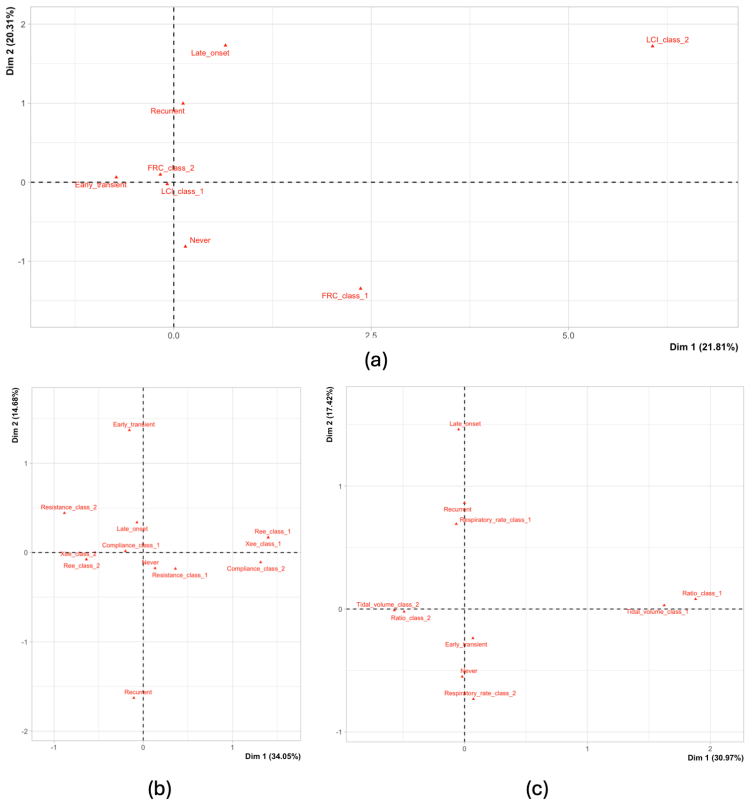


Figure 5.1: MCA biplots for (a) MBW, (b) airwave oscillometry, and (c) tidal breathing measures

Each of these lung function outcomes captures different aspects of respiratory physiology, oscillometry measures airway resistance and reactance, tidal breathing reflects general ventilation efficiency, and MBW assesses ventilation distribution and gas mixing. In some parts of the analysis, these outcomes tend to form distinct clusters or associations with particular wheeze phenotypes. An example of this in Figure 5.1 (c) is the close proximity of the recurrent wheezing and respiratory rate class 1 categories. Generally speaking, the points that are close to one another in the plot indicate that the categories are associated. Another important observation is that the MCA plots illustrate

how different classes of lung functions are grouped together. For example, in Figure 5.1 (c), tidal volume class 1 and t_{PTEF}/t_E class 1, both representing profiles with higher baselines that decrease over time, are positioned close to each other. While tidal volume class 2 and t_{PTEF}/t_E class 2 are located together on the opposite end of dimension 1.

Figure 5.1 does seem to suggest some separation between the variables along the two dimensions. When the points align along the axes with little overlap, it indicates that the first dimension and the second dimension are explaining different sources of variability in the data. Although this study has provided evidence that there are associations between lung function outcomes and wheezing phenotypes, the results presented in Chapter 5 enforce the idea that lung function outcomes and wheezing patterns are influenced by different underlying factors in the cohort, and each contributes uniquely to the overall structure of the data. Respiratory health is shaped by multiple, potentially separate processes. These include changes in the structure of the lungs, such as airway development, as well as functional changes, like airflow or gas exchange efficiency. These processes may evolve alongside wheezing, but they could be driven by different mechanisms. This highlights the need to consider both structural and functional aspects when interpreting lung function and wheezing patterns. The preceding work found that lung function outcomes and wheezing patterns are influenced by different underlying factors. This separation in the influence of predictors aligns with the idea that different underlying factors drive the observed patterns in lung function and wheezing, supporting the notion of distinct contributions to the data's structure.

The final integration using MCA visually represented the associations between wheezing phenotypes and lung function outcomes, illustrating how distinct respiratory trajectories may reflect separate underlying processes. Overall, this research underscores the complex and interconnected factors that shape early respiratory health, highlighting the need for further study into the mechanisms influencing lung function and wheezing patterns in diverse populations.

Appendix A

A.1 Chapter 3

Table A.1: Comparison of included versus excluded children in the DCHS

	Included (N=950)	Excluded or Lost to Follow-Up (N=193)	P-value
Maternal characteristics			
Median age [IQR] at enrolment (years)	25.8 [22.0-30.8]	24.4 [21.7-30.2]	0.061
Antenatal smoking	286/950 (31.3%)	38/193 (19.7%)	< 0.0001
Maternal allergy	55/879 (6.3%)	10/182 (5.5%)	0.82
Antenatal depression	203/845 (24.0%)	34/149 (22.8%)	0.97
Antenatal psychological distress	174/846 (20.6%)	27/145 (18.6%)	0.67
Antenatal IPV	292/849 (34.4%)	42/150 (28.0%)	0.15
Mode of delivery (Caesarean)	188/950 (19.8%)	41/193 (21.2%)	0.72
Infant characteristics			
Sex (male)	481/950 (50.6%)	105/193 (54.4%)	0.38
Pre-term (<37 weeks)	150/950 (15.8%)	41/193 (21.2%)	0.081
Late preterm (34 to <37 weeks)	101/950 (10.6%)	22/193 (11.4%)	0.85
HIV exposed uninfected	206/950 (21.7%)	42/193 (21.7%)	0.99
Exclusive Breastfeeding at 6 weeks	468/950 (49.3%)	52/175 (29.7%)	< 0.0001
Season of birth			
Summer	241/950 (25.4%)	46/193 (23.8%)	0.71
Autumn	239/950 (25.2%)	55/193 (28.5%)	0.37
Winter	256/950 (26.9%)	50/193 (25.9%)	0.83
Spring	214/950 (22.5%)	42/193 (21.8%)	0.89
Median weight-for-age z-score at birth [IQR]	-0.57 [-1.33; 0.09]	-0.55 [-1.21; -0.04]	0.87
Socioeconomic status (SES)			
Income			
<ZAR1 000 ($67)	374/950 (39.4%)	57/193 (29.5%)	0.013
1 000-5 000 ($67-$336)	462/950 (48.6%)	95/193 (49.2%)	0.94
>ZAR5 000 ($336)	114/950 (12.0%)	41/193 (21.2%)	0.00095
Asset ownership			
Low	242/950 (25.5%)	55/193 (28.5%)	0.43
Low-Medium	296/950 (31.2%)	47/193 (24.3%)	0.073
Medium-High	221/950 (23.3%)	40/193 (20.7%)	0.51
High	191/950 (20.1%)	51/193 (26.4%)	0.063
Household size			
Small [1-4)	302/948 (31.8%)	79/193 (40.9%)	0.018
Small-Medium [4-5)	182/948 (19.2%)	32/193 (16.6%)	0.45
Medium-Large [5-7)	259/948 (27.3%)	48/193 (24.9%)	0.54
Large [7-18]	205/948 (21.6%)	34/193 (17.6%)	0.25
Education			
Primary	70/950 (7.4%)	16/193 (8.3%)	0.76
Some secondary	523/950 (55.1%)	86/193 (44.6%)	0.0097
Completed secondary	303/950 (31.9%)	72/193 (37.3%)	0.17
Any tertiary	54/950 (5.7%)	19/193 (9.8%)	0.046
Lung function (oscillometry) at 6-weeks			
Ree ($hPa.s.L^{-1}$)	42.8 [37.1-50.9]	40.8 [33.5-50.4]	0.68
Xee ($hPa.s.L^{-1}$)	-6.7 [-11.9; -2.9]	-6.4 [-10.8; -2.2]	0.51

LRTI = Lower Respiratory Tract Infection; IPV = intimate partner violence; Ree = Respiratory resistance at the end of expiration
Xee = Respiratory reactance at the end of expiration

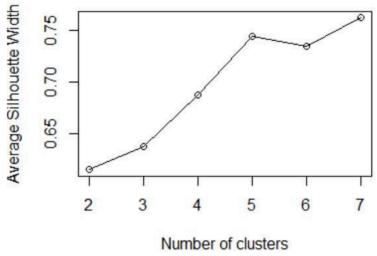


Figure A.1: Plot of average silhouette width in ALSPAC to determine optimal number of clusters using the PAM algorithm.

Table A.2: Bayesian Information Criterion (BIC) values for different phenotype sizes which were used by the LCA model to identify the wheezing phenotypes in Figure 3.9.

Number of Phenotypes	BIC
2	5954.243
3	5998.755
4	6068.423
5	6146.388
6	6218.691
7	6288.906

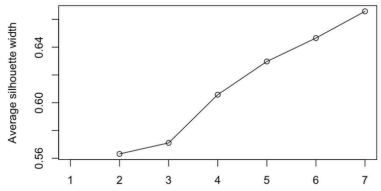
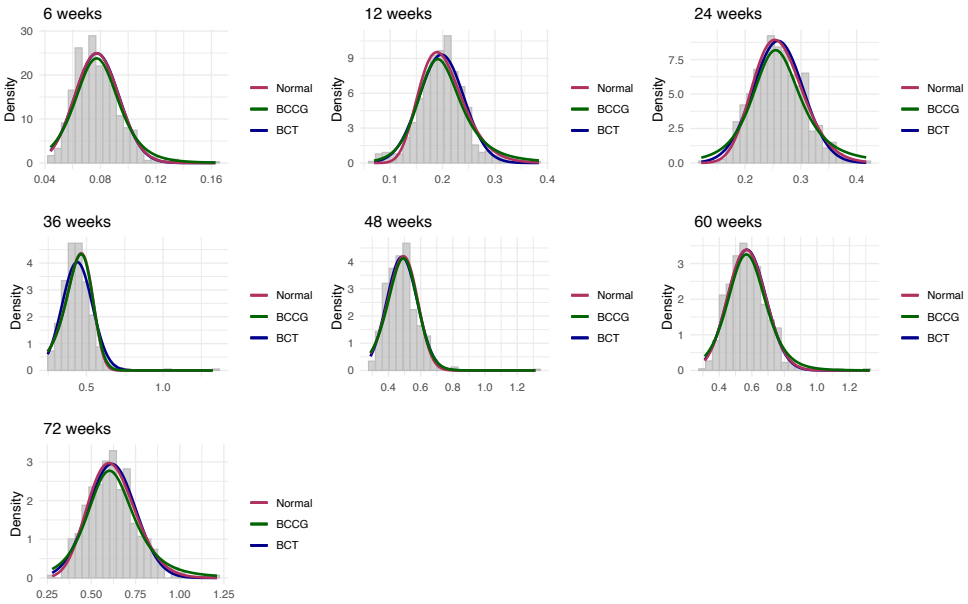
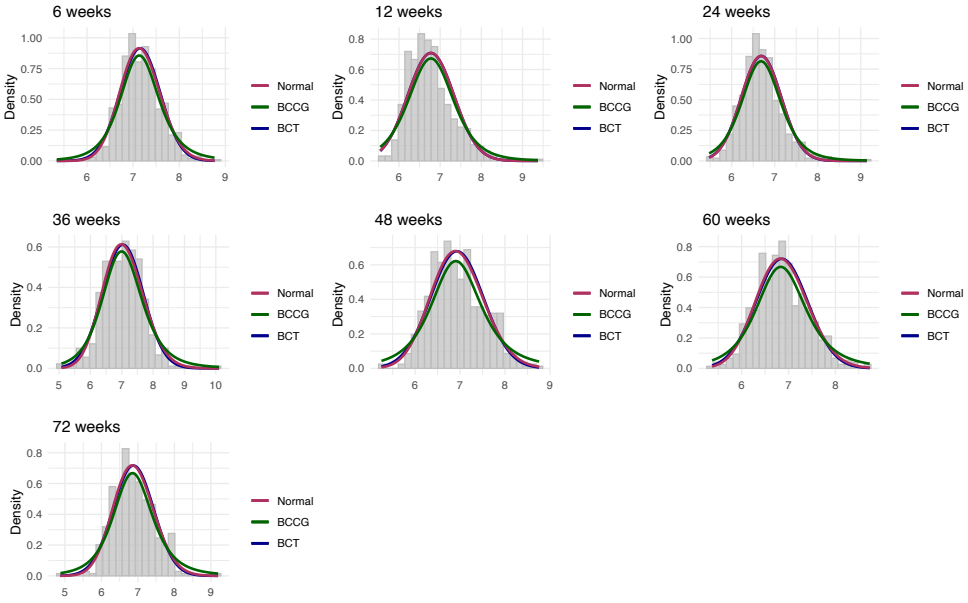


Figure A.2: Plot of average silhouette width to determine optimal number of clusters using the PAM algorithm - application of PAM to binary wheezing outcomes

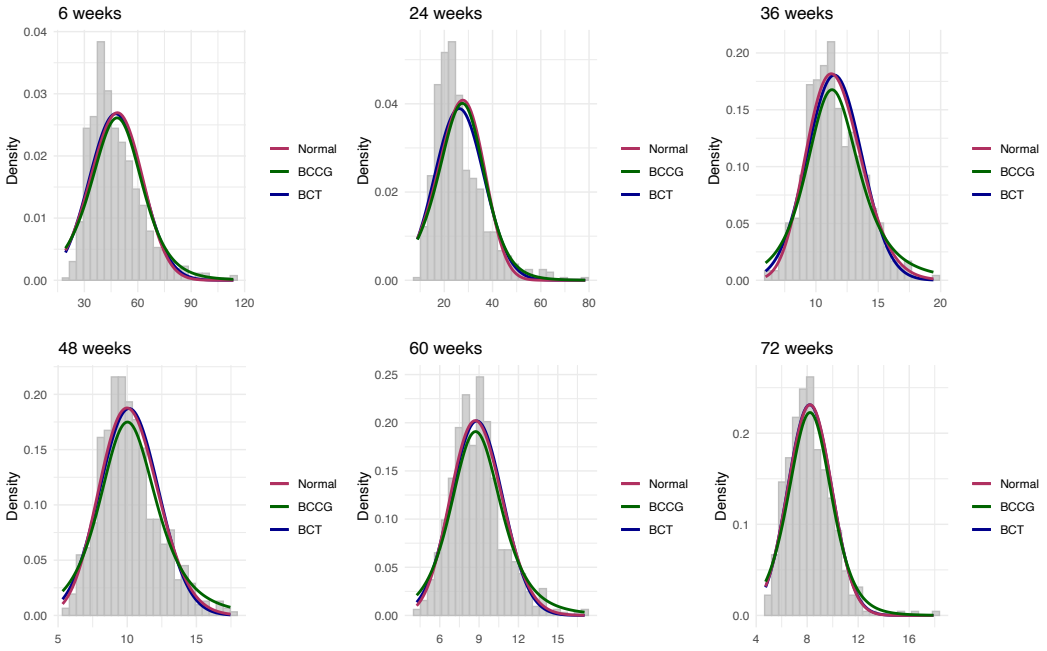
A.2 Chapter 4



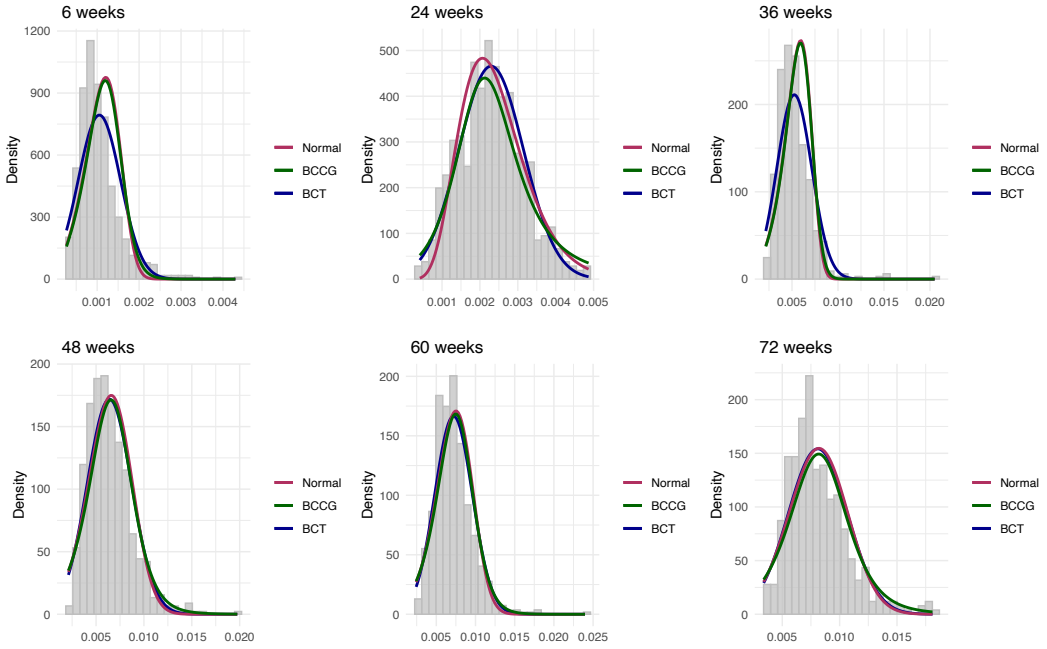
(a)



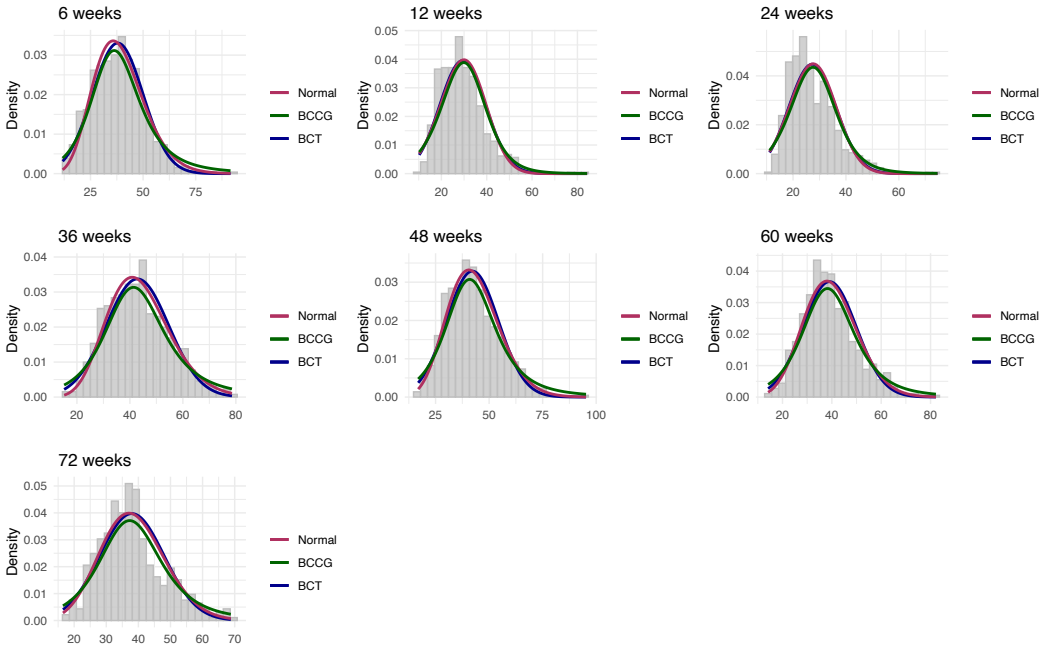
(b)


6 weeks
24 weeks
36 weeks
48 weeks
60 weeks
72 weeks
Density
Normal
BCCG
BCT

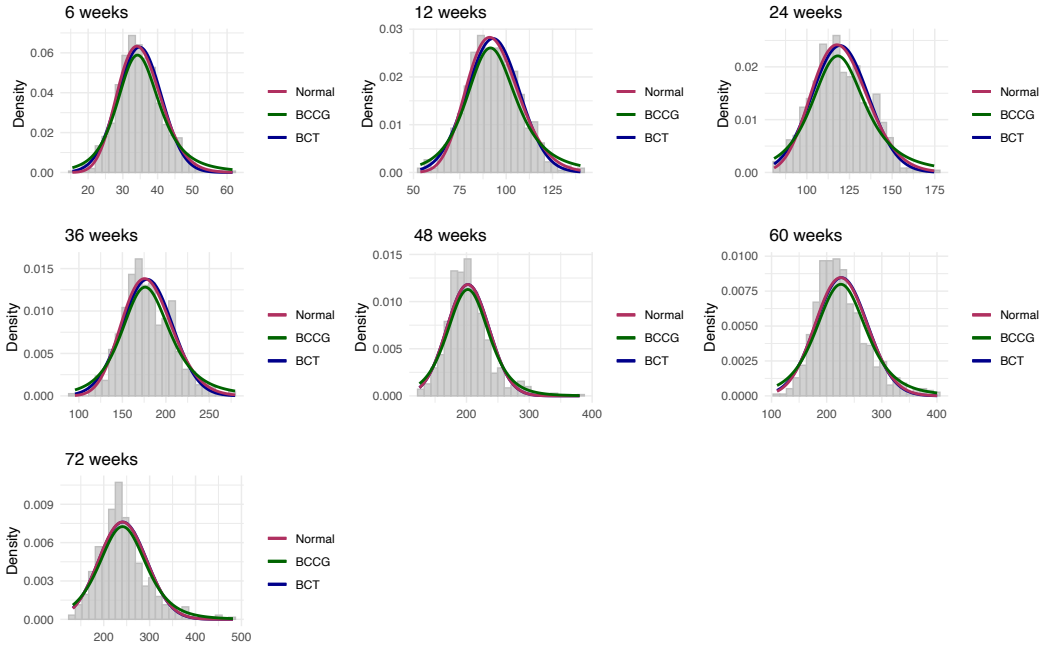
(c)


6 weeks
24 weeks
36 weeks
48 weeks
60 weeks
72 weeks
Density
Normal
BCCG
BCT

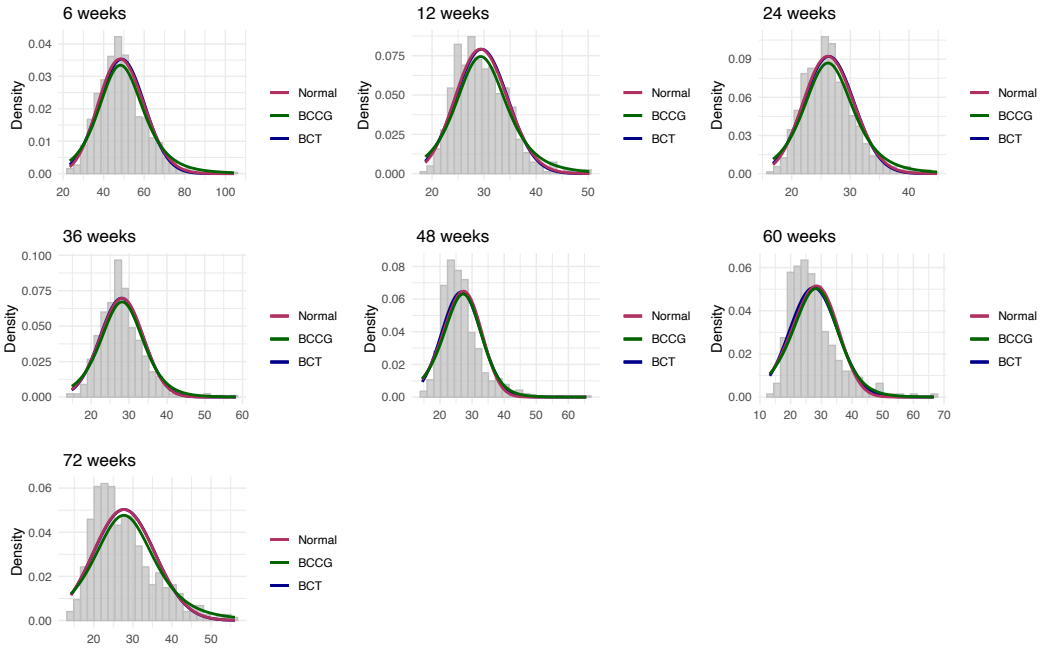
(d)


6 weeks
12 weeks
24 weeks
36 weeks
48 weeks
60 weeks
72 weeks
Density
Normal
BCCG
BCT

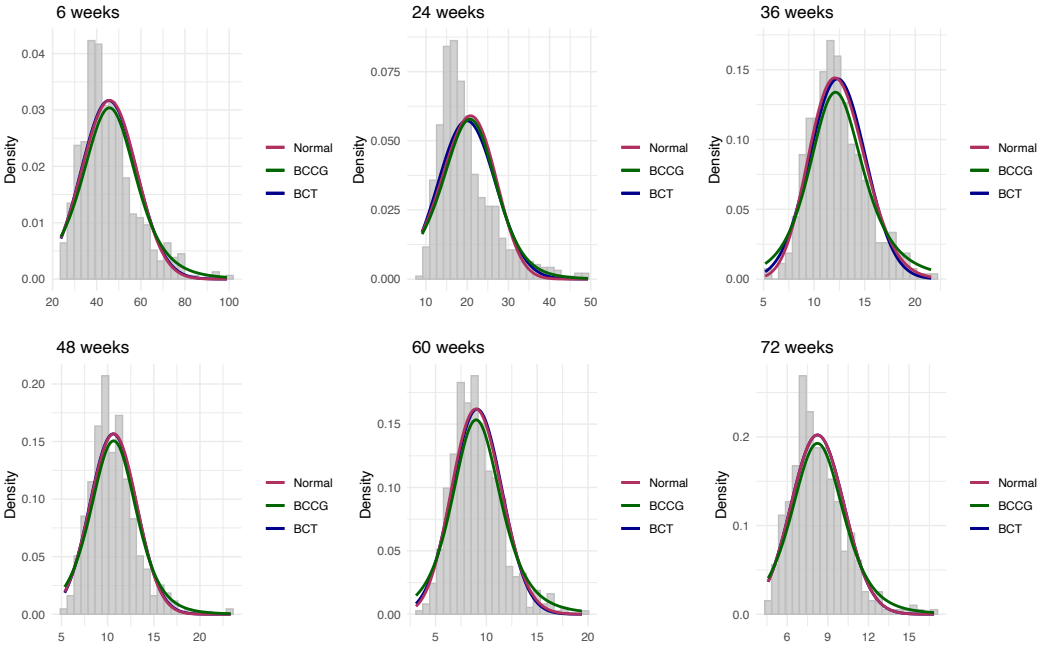
(e)


6 weeks
12 weeks
24 weeks
36 weeks
48 weeks
60 weeks
72 weeks
Density
Normal
BCCG
BCT

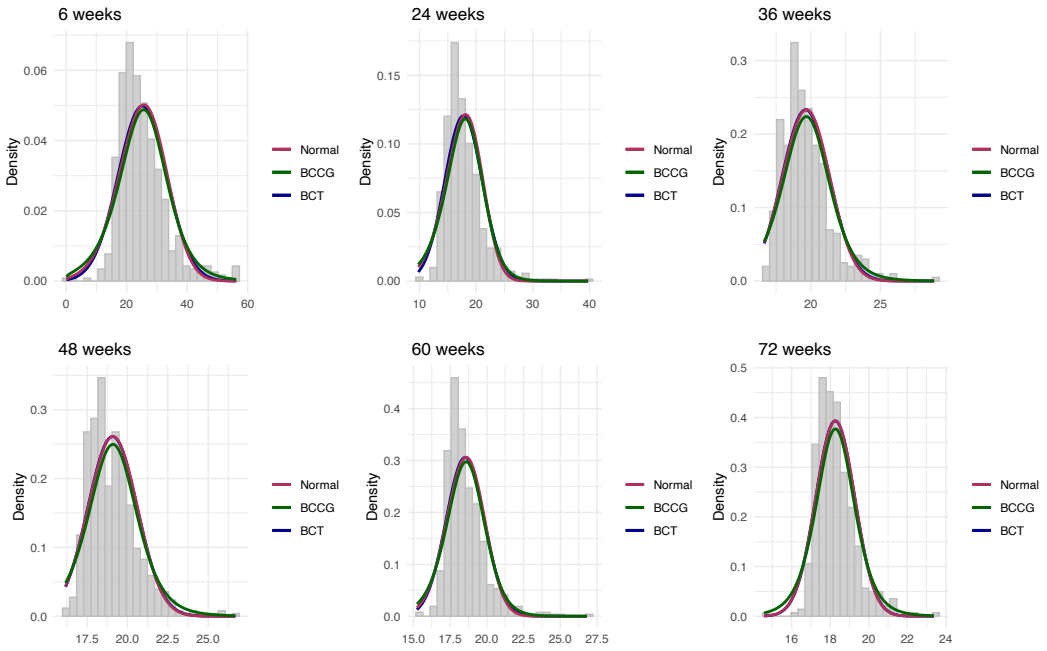
(f)


6 weeks
12 weeks
24 weeks
36 weeks
48 weeks
60 weeks
72 weeks
Density
Normal
BCCG
BCT

(g)


6 weeks
24 weeks
36 weeks
48 weeks
60 weeks
72 weeks
Density
Normal
BCCG
BCT

(h)



(i)

Figure A.2: Density histograms with theoretical densities of normal, BCCG, BCT distributions overlaid for each measured lung function outcome (a) functional residual capacity, (b) lung clearance index, (c) resistance, (d) compliance, (e) Ratio time of peak total expiratory flow to time of expiration (t_{PTEF}/t_E), (f) tidal volume, (g) respiratory resistance at the end of expiration (Ree), and (g) respiratory reactance at the end of expiration (Xee).

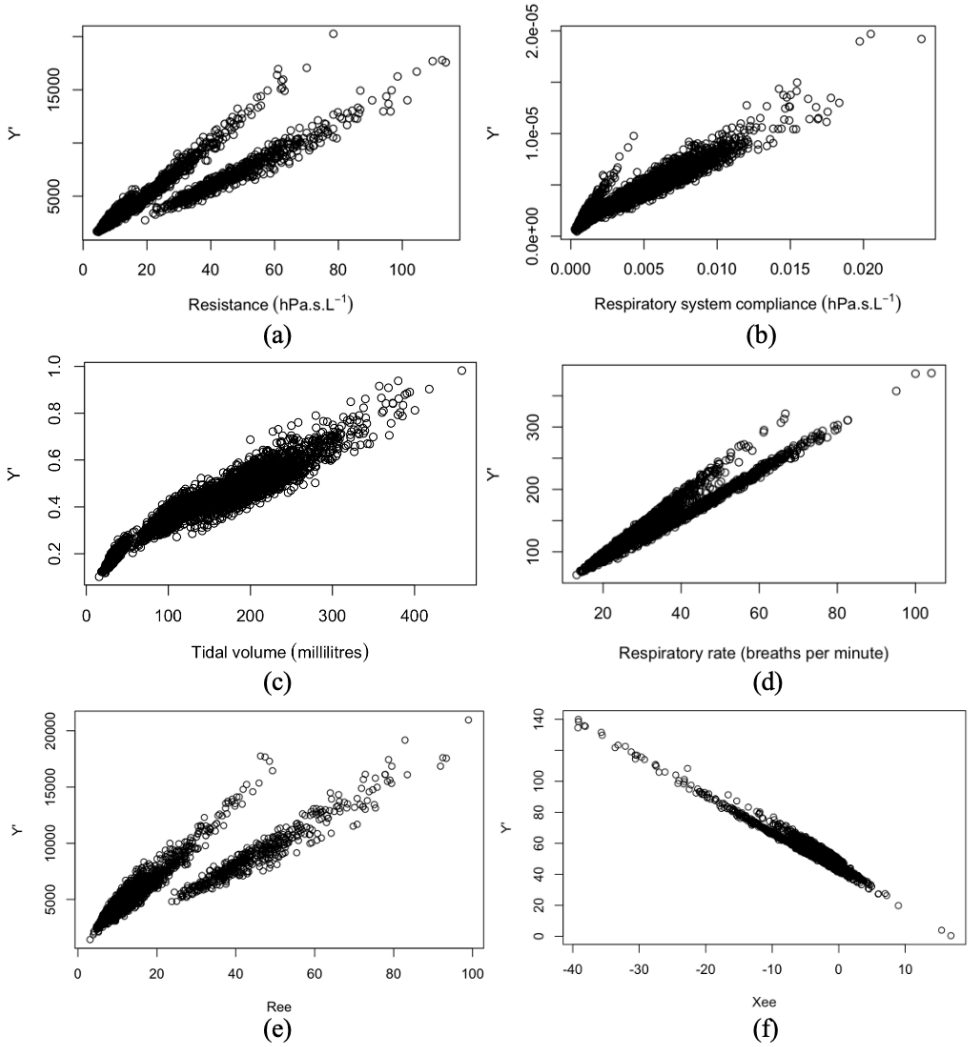


Figure A.3: Unstandardised lung function outcomes plotted against the height-adjusted outcomes (Y') for (a) resistance, (b) compliance, (c) tidal volume, (d) respiratory rate, (e) respiratory resistance at the end of expiration (Ree), and (f) respiratory reactance at the end of expiration (Xee)

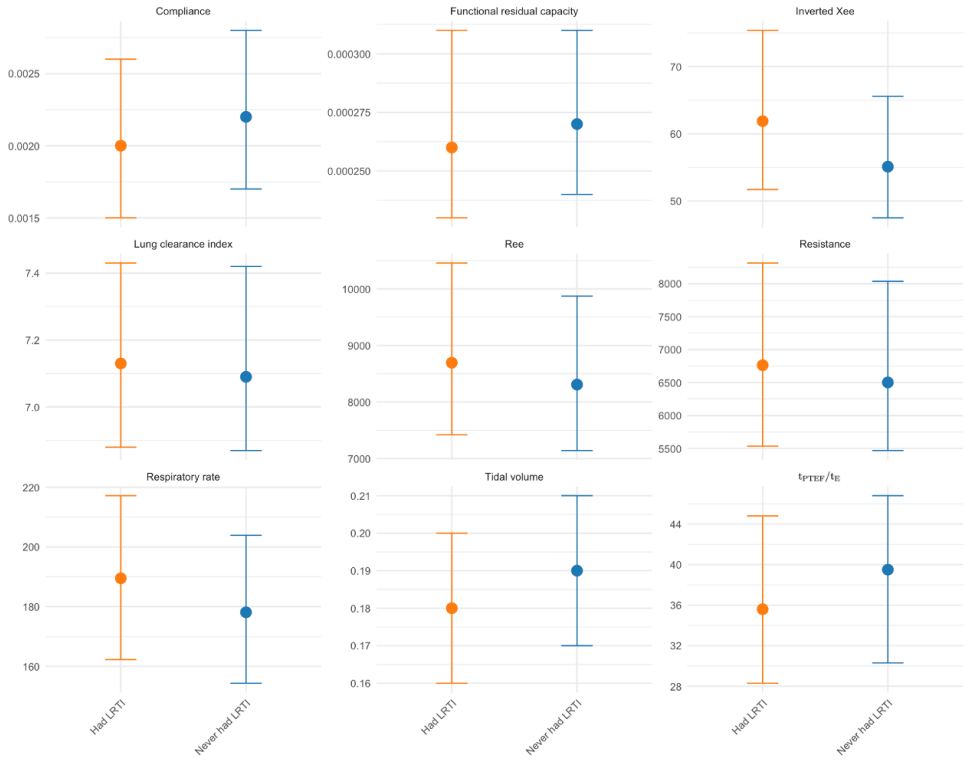


Figure A.4: Lung function outcomes by LRTI history: never versus ever diagnosed. A visual representation of the medians with interquartile ranges (IQR) shown in Table 4.13

	FRC[1]		LCI[1]		Respiratory rate[1]		Tidal volume[2]		t_{PTEF}/t_E[1]	
	Effect size (95% CI)	P-value (adjusted for multiple comparisons[3])	Effect size (95% CI)	P-value (adjusted for multiple comparisons[3])	Effect size (95% CI)	P-value (adjusted for multiple comparisons[3])	Effect size (95% CI)	P-value (adjusted for multiple comparisons[3])	Effect size (95% CI)	P-value (adjusted for multiple comparisons[3])
LRTI ever	0.987 (0.975-0.998)	0.022 (0.055)	0.999 (0.994-1.005)	0.91 (0.97)	1.018 (1.0063-1.029)	0.0024 (0.0067)	0.993 (0.984-1.003)	0.18 (0.31)	1.002 (0.984-1.020)	0.79 (0.82)
LRTI ever under 1 year (versus no LRTI)	0.990 (0.979-1.002)	0.11 (0.20)	1.001 (0.995-1.006)	0.79 (0.94)	1.025 (1.013-1.037)	<0.0001 (<0.0001)	0.986 (0.977-0.996)	0.0059 (0.013)	1.005 (0.987-1.023)	0.59 (0.69)
LRTI count birth to 5 years	0.997 (0.9931-1.001)	0.12 (0.21)	1.002 (1.0003-1.004)	0.019 (0.095)	1.012 (1.009-1.016)	<0.0001 (<0.0001)	0.996 (0.993-0.998)	0.0081 (0.018)	0.999 (0.994-1.006)	0.92 (0.92)
Hospitalised LRTI (vs ambulatory LRTI)	1.010 (0.995-1.026)	0.21 (0.29)	1.002 (0.995-1.009)	0.58 (0.85)	0.979 (0.964-0.994)	0.0054 (0.015)	0.997 (0.984-1.010)	0.64 (0.73)	1.001 (0.977-1.024)	0.95 (0.95)
RSV-hospitalised LRTI (vs RSV-ambulatory LRTI)	0.989 (0.966-1.013)	0.39 (0.51)	0.998 (0.988-1.010)	0.82 (0.95)	0.955 (0.933-0.978)	<0.0001 (<0.0001)	1.003 (0.983-1.023)	0.76 (0.86)	0.985 (0.949-1.023)	0.43 (0.54)
RV-hospitalised LRTI (vs RV-ambulatory LRTI)	1.008 (0.984-1.032)	0.51 (0.64)	0.994 (0.983-1.005)	0.28 (0.72)	0.967 (0.944-0.989)	0.0051 (0.014)	1.019 (0.999-1.039)	0.062 (0.14)	1.001 (0.965-1.039)	0.95 (0.95)
RSV-LRTI (vs non-RSV LRTI)	1.014 (0.998-1.029)	0.076 (0.15)	0.999 (0.993-1.007)	0.98 (0.98)	0.992 (0.977-1.007)	0.29 (0.40)	1.013 (0.999-1.0252)	0.051 (0.012)	0.971 (0.949-0.994)	0.014 (0.032)
RV-LRTI (vs non-RV LRTI)	1.008 (0.993-1.022)	0.13 (0.22)	1.005 (0.991-1.020)	0.95 (0.97)	0.999 (0.993-1.007)	0.31 (0.43)	1.004 (0.992-1.016)	0.49 (0.56)	1.018 (0.995-1.041)	0.51 (0.58)

Table notes:
All models include a child-specific random effect.
[1]Box-Cox-t (BCT) distribution.
[2]Box-Cox Cole and Green (BCCG) distribution.
[3] An approach by Benjamini, Y., and Hochberg, Y. (1995) for multiple testing.
FRC = functional residual capacity; LCI = lung clearance index; t_{PTEF}/t_E = time to peak tidal expiratory time over total expiratory time; IPV = intimate partner violence, LRTI = lower respiratory tract infection, RSV = respiratory syncytial virus; RV = Rhinovirus
The models for LCI and t_{PTEF}/t_E were done using the unstandardised measurements
Effect size: exponent of the μ-coefficients

Table A.3: Association of specific LRTI with the development of multiple breath washout and tidal breathing parameters from birth to 5 years

	Resistance[1]		Compliance[1]		Ree[1]		Inversed Xee[2]	
	Effect size (95% CI)	P-value (adjusted for multiple comparisons[3])	Effect size (95% CI)	P-value (adjusted for multiple comparisons[3])	Effect size (95% CI)	P-value (adjusted for multiple comparisons[3])	Effect size (95% CI)	P-value (adjusted for multiple comparisons[3])
LRTI ever	1.028 (1.016-1.041)	<0.0001 (<0.0001)	0.959 (0.941-0.978)	<0.0001 (<0.0001)	1.034 (1.017-1.051)	<0.0001 (<0.0001)	1.0059 (1.0015-1.0104)	0.0083 (0.029)
LRTI ever under 1 year (versus no LRTI)	1.035 (1.022-1.048)	<0.0001 (<0.0001)	0.958 (0.938-0.978)	<0.0001 (<0.0001)	1.037 (1.020-1.054)	<0.00001 (<0.0001)	1.011 (1.006-1.015)	<0.0001 (<0.0001)
LRTI count birth to 5 years	1.002 (0.998-1.006)	0.26 (0.54)	0.987 (0.980-0.993)	<0.0001 (<0.0001)	1.007 (1.002-1.013)	0.0079 (0.025)	1.003 (1.001-1.004)	<0.0001 (<0.0001)
Hospitalised LRTI (vs ambulatory LRTI)	0.996 (0.980-1.013)	0.67 (0.79)	0.988 (0.963-1.015)	0.39 (0.46)	1.001 (0.980-1.022)	0.91 (0.92)	1.007 (1.001-1.013)	0.021 (0.052)
RSV-LRTI hospitalised (vs RSV-ambulatory LRTI)	1.020 (0.994-1.047)	0.13 (0.29)	0.938 (0.899-0.978)	0.0030 (0.0080)	1.022 (0.988-1.057)	0.21 (0.33)	1.009 (1.0001-1.019)	0.048 (0.10)
RV-LRTI hospitalised (vs RV-ambulatory LRTI)	1.005 (0.979-1.032)	0.69 (0.82)	0.985 (0.944-1.027)	0.48 (0.54)	1.006 (0.973-1.041)	0.71 (0.89)	1.002 (0.993-1.012)	0.64 (0.84)
RSV-LRTI (vs non-RSV LRTI)	1.011 (0.994-1.027)	0.21 (0.43)	0.969 (0.945-0.995)	0.022 (0.055)	1.017 (0.995-1.038)	0.12 (0.23)	1.014 (1.008-1.019)	<0.0001 (<0.0001)
RV-LRTI (vs non-RV LRTI)	1.002 (0.986-1.018)	0.79 (0.86)	0.989 (0.965-1.015)	0.43 (0.11)	1.013 (0.992-1.033)	0.23 (0.36)	1.003 (0.997-1.009)	0.28 (0.43)

Table notes:
All models include a child-specific random effect.
[1]Box-Cox-t (BCT) distribution.
[2]Box-Cox Cole and Green (BCCG) distribution.
[3] An approach by Benjamini, Y., and Hochberg, Y. (1995) for multiple testing.
Ree = respiratory resistance at the end of expiration; Xee = respiratory reactance at the end of expiration; IPV = intimate partner violence, LRTI = lower respiratory tract infection, RSV = respiratory syncytial virus; RV = Rhinovirus
Effect size: exponent of the μ-coefficients

Table A.4: Association of LRTIs with the development of oscillometry parameters from birth to age 5 years.

A.2.1 Multivariate LCMM outcomes for oscillometry outcomes

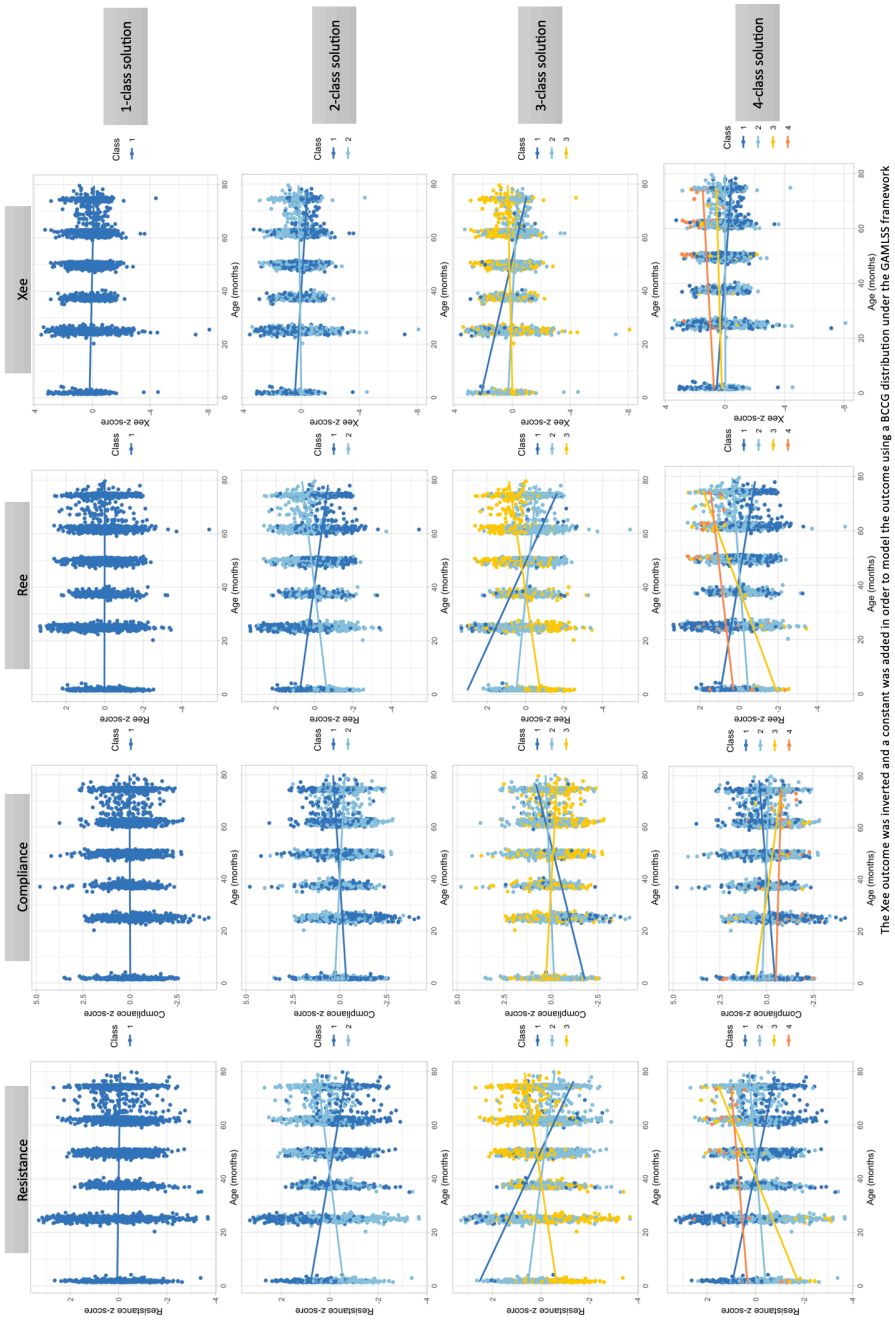


Figure A.5: This figure illustrates the heterogeneous structures within longitudinal multivariate airwave oscillometry outcomes using linear models. Points show individual scores, while classes were computed through the loess function in R for visualisation purposes.

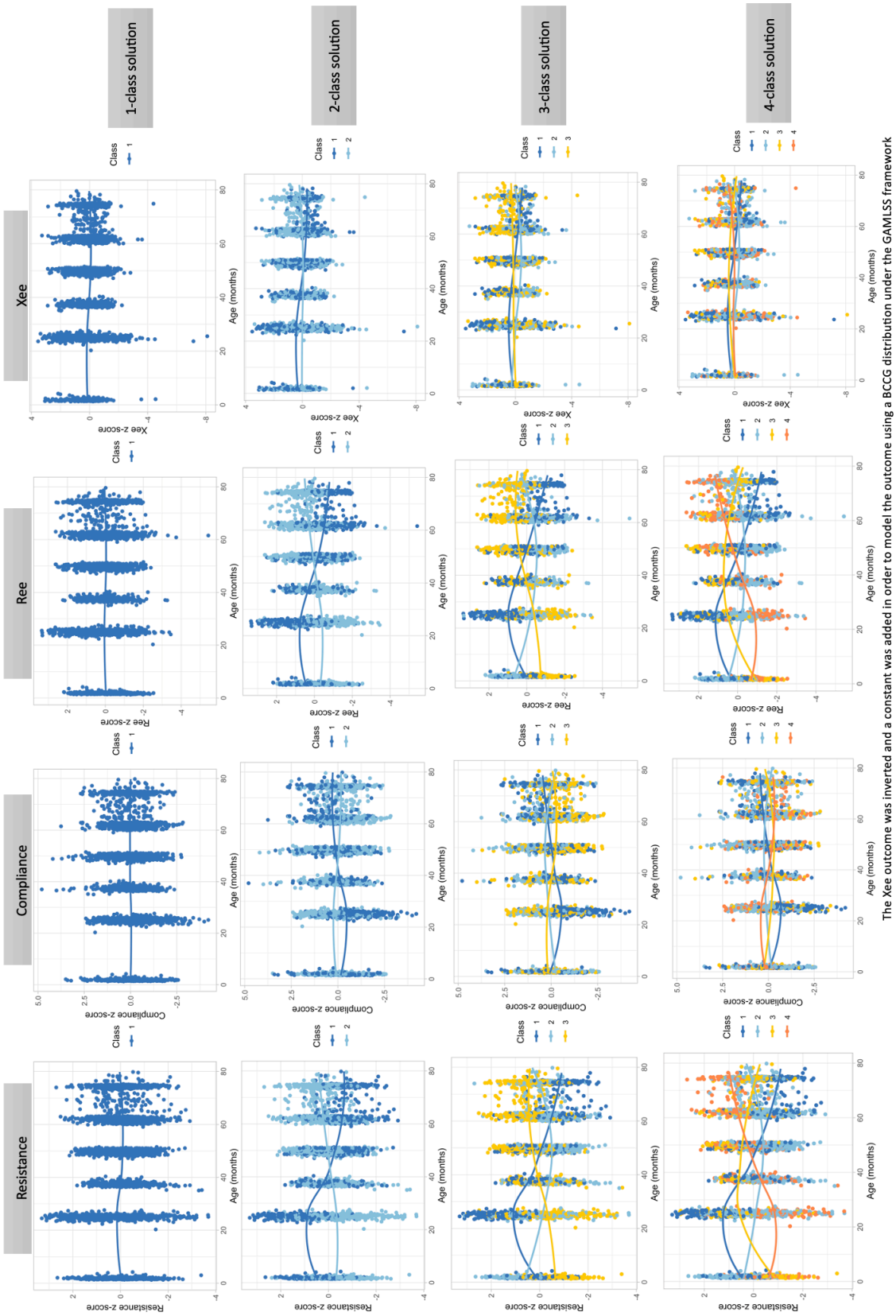


Figure A.6: This figure illustrates the heterogeneous structures within longitudinal multivariate airwave oscillometry outcomes using quadratic models. Points show individual scores, while classes were computed through the loess function in R for visualisation purposes.

Table A.5: Summary table: Multivariate airwave oscillometry linear and quadratic models depicted in Figure A.5 and A.6

Variable	**G**	**loglik**	**npm**	**AIC**	**BIC**	**entropy**	**Class 1 (%)**	**Class 2 (%)**	**Class 3 (%)**	**Class 4 (%)**
Linear	1	-17694.70	15	35419.39	35492.18	1.00	100.00%			
	2	-29127.81	18	58291.62	58378.96	0.45	48.10%	51.90%		
	3	-29103.22	21	58248.44	58350.34	0.59	3.38%	55.18%	41.44%	
	4	-29122.21	24	58292.42	58408.88	0.57	36.89%	58.03%	2.96%	2.11%
Quadratic	1	-17693.40	16	35418.81	35496.44	1.00	100.00%			
	2	-29109.46	20	58258.92	58355.96	0.48	38.37%	61.63%		
	3	-29036.93	24	58121.86	58238.32	0.47	25.69%	35.41%	38.90%	
	4	-29005.29	28	58066.58	58202.45	0.49	17.23%	47.99%	18.39%	16.38%

A.2.2 Multivariate LCMM outcomes for MBW and tidal breathing measures

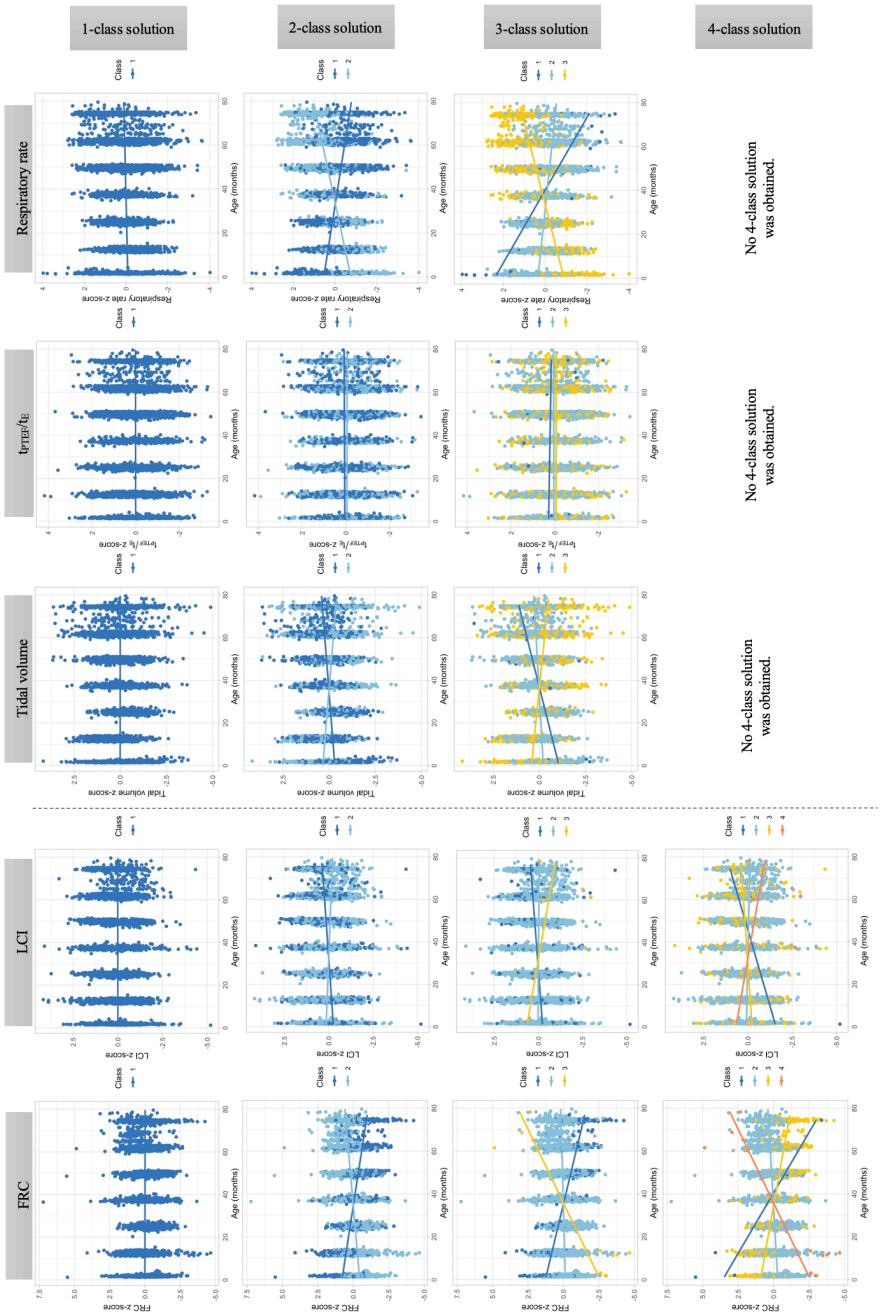


Figure A.7: This figure illustrates the heterogeneous structures within longitudinal multivariate MBW (left) and tidal breathing (right) outcomes using linear models. Points show individual scores, while classes were computed through the loess function in R for visualisation purposes.

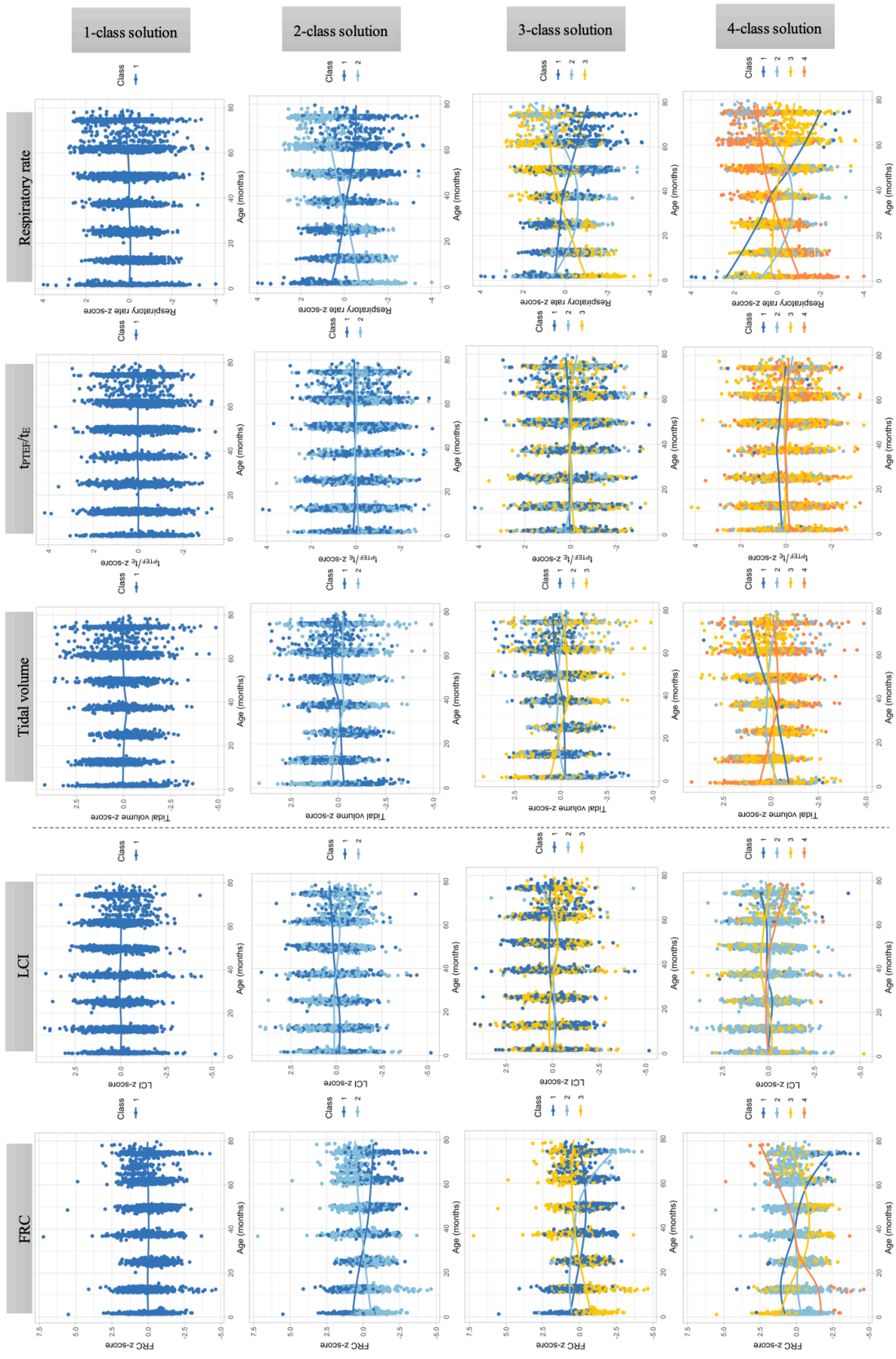


Figure A.8: This figure illustrates the heterogeneous structures within longitudinal multivariate MBW (left) and tidal breathing (right) outcomes using quadratic models. Points show individual scores, while classes were computed through the loess function in R for visualisation purposes.

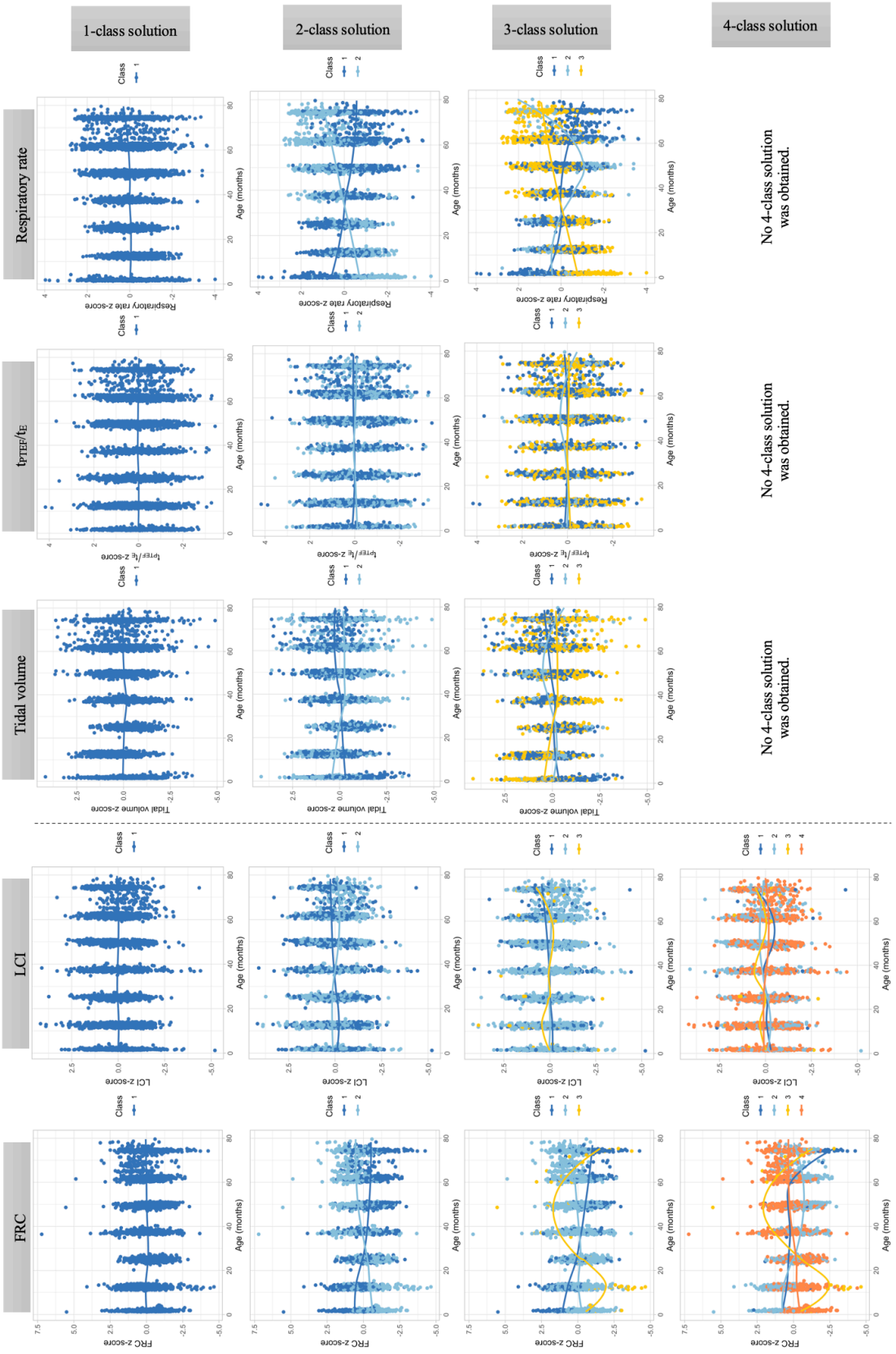


Figure A.9: This figure illustrates the heterogeneous structures within longitudinal multivariate MBW (left) and tidal breathing (right) outcomes using spline models. Points show individual scores, while classes were computed through the loess function in R for visualisation purposes.

Table A.6: Summary table: Multivariate MBW and tidal breathing linear, quadratic, and spline models depicted in Figure A.7, A.8, and A.9

Variable	G	loglik	npm	AIC	BIC	entropy	Class 1 (%)	Class 2 (%)	Class 3 (%)	Class 4 (%)
					MBW outcomes					
Linear	1	-10874.72	9	21767.43	21811.23	1.00	100%			
	2	-17894.35	12	35812.71	35871.10	0.23	35.25%	64.75%		
	3	-17888.93	15	35807.86	35880.85	0.61	10.22%	88.84%	0.94%	
	4	-17883.16	18	35802.31	35889.90	0.60	0.31%	75.39%	23.36%	0.94%
Quadratic	1	-10874.21	10	21768.42	21817.08	1.00	100%			
	2	-17894.01	14	35816.01	35884.14	0.22	41.71%	58.29%		
	3	-17886.24	18	35808.48	35896.06	0.35	55.89%	2.92%	41.19%	
	4	-17881.49	22	35806.98	35914.03	0.44	3.02%	76.02%	18.98 %	1.98%
Spline	1	-10872.47	11	21766.95	21820.47	1.00	100%			
	2	-17893.48	16	35818.95	35896.81	0.21	50.16%	49.84%		
	3	-17875.43	21	35792.85	35895.04	0.49	19.19%	78.21%	2.61 %	
	4	-17861.69	26	35775.38	35901.89	0.52	2.09%	32.12%	1.36%	64.44%
					Tidal breathing outcomes					
Linear	1	-18106.80	12	36237.59	36296.04	1.00	100.00%			
	2	-29754.04	15	59538.08	59611.15	0.39	53.42%	46.58%		
	3	-29743.08	18	59522.16	59609.84	0.58	1.56%	63.07%	35.37%	
	4	-29751.31	21	59544.62	59646.91	0.69	51.24%	48.65%	0.10 %	0.00%
Quadratic	1	-18098.87	13	36223.73	36287.05	1.00	100.00%			
	2	-29750.93	17	59535.86	59618.67	0.38	53.63%	46.37%		
	3	-29734.01	21	59510.02	59612.31	0.35	43.67	21.89%	34.44%	
	4	-29721.78	25	59493.56	59615.34	0.47	1.56%	16.70%	50.73%	31.02%
Spline	1	-18099.78	14	36227.55	36295.75	1.00	100.00%			
	2	-29747.00	19	59532.01	59624.56	0.39	53.32%	46.68%		
	3	-29721.89	24	59491.77	59608.68	0.40	45.75%	10.99%	43.26%	
	4	-29723.30	29	59504.59	59645.86	0.52	11.10%	42.84%	46.06%	0.00%

A.2.3 Univariate LCMM outcomes for oscillometry outcomes

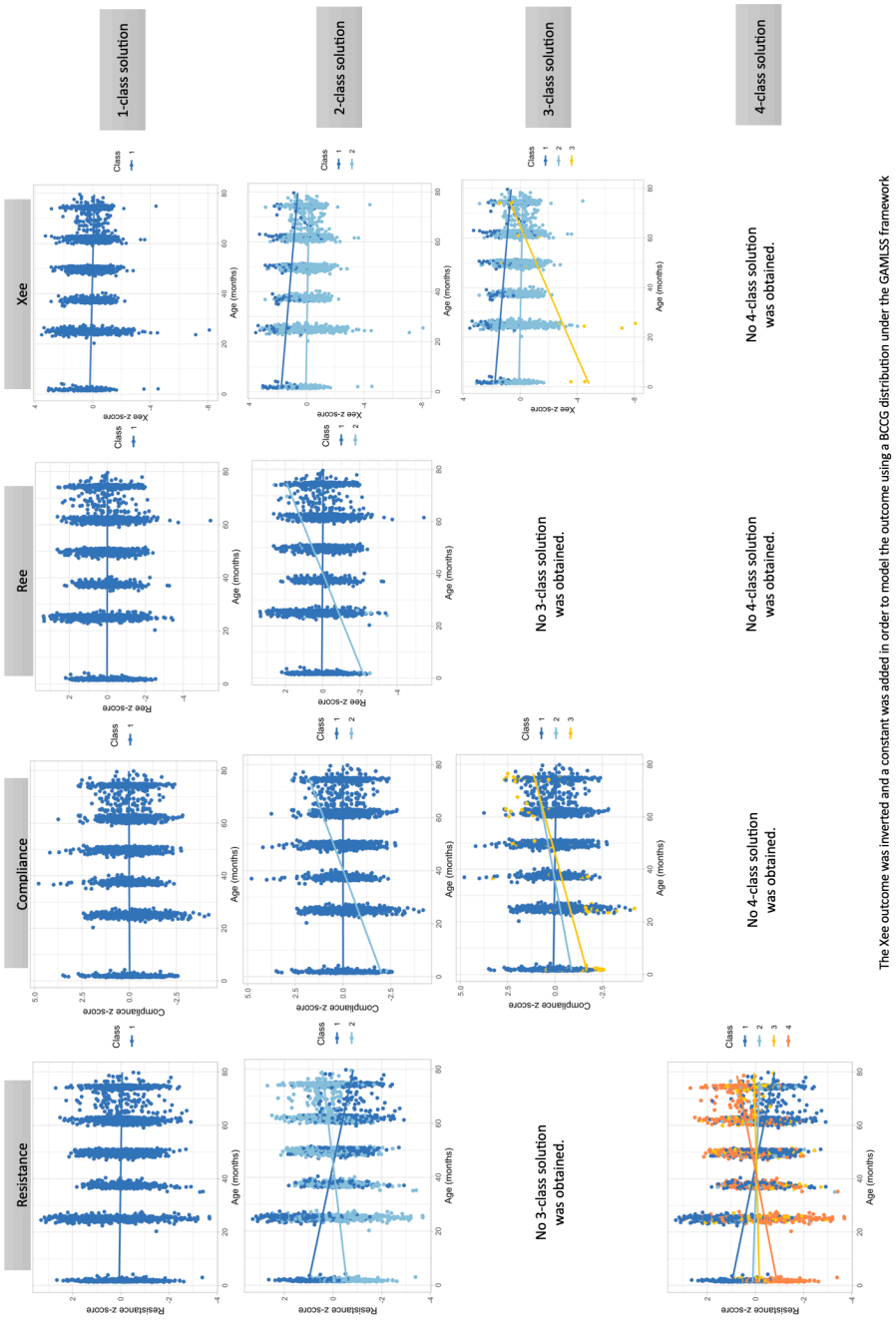


Figure A.10: This figure illustrates the heterogeneous structures within longitudinal univariate airwave oscillometry outcomes using linear models. Points show individual scores, while classes were computed through the loess function in R for visualisation purposes.

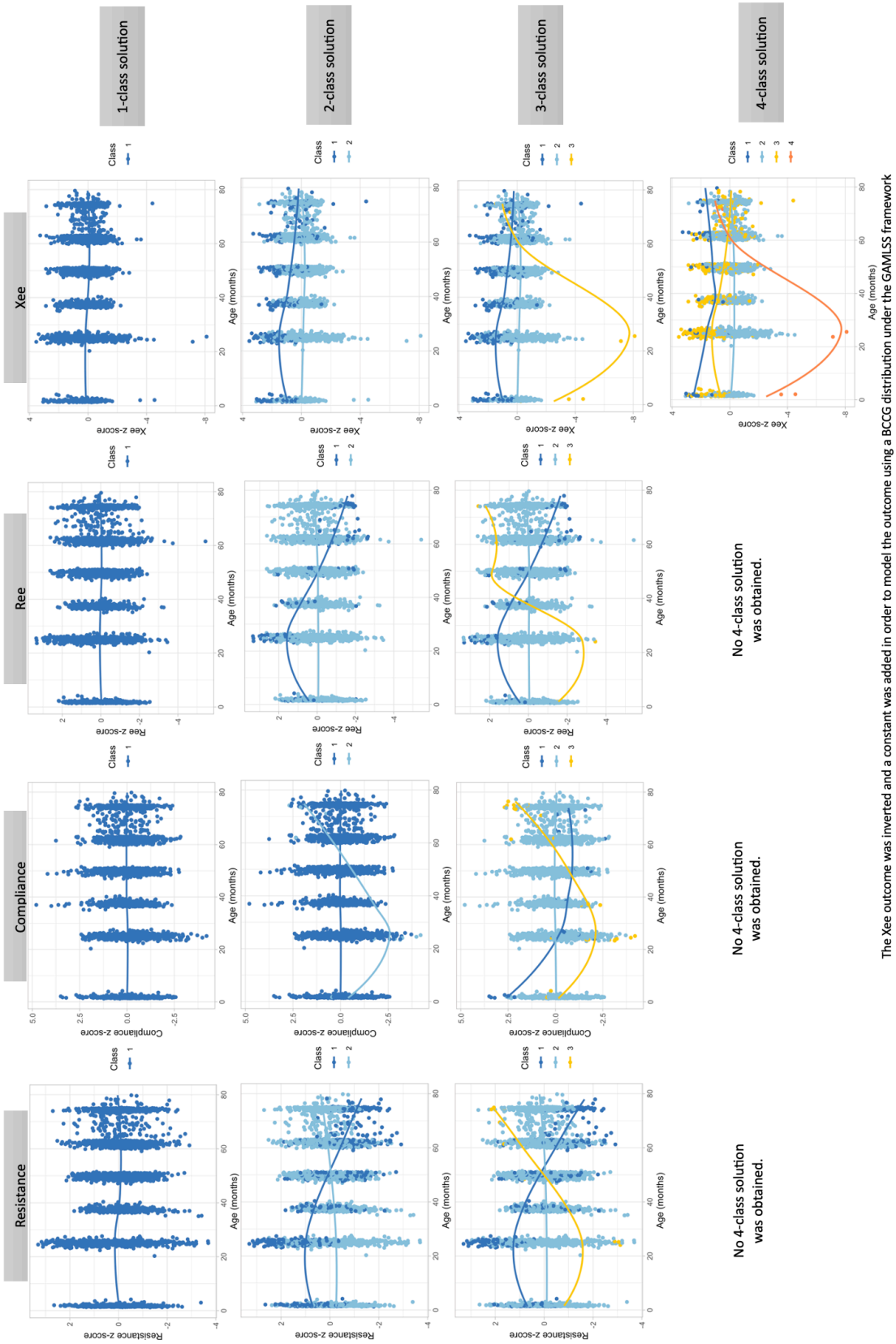


Figure A.11: This figure illustrates the heterogeneous structures within longitudinal univariate airwave oscillometry outcomes using quadratic models. Points show individual scores, while classes were computed through the loess function in R for visualisation purposes.

Table A.7: Summary table: Univariate airwave oscillometry linear models depicted in Figure A.10

Variable	G	loglik	npm	AIC	BIC	entropy	Class 1 (%)	Class 2 (%)	Class 3 (%)	Class 4 (%)
Resistance	1	-4672.56	6	9357.13	9386.24	1.00	100%			
	2	-4672.39	9	9362.78	9406.45	0.11	40.06%	59.94%		
	3	-4672.39	12	9368.78	9427.01	0.09	65.01%	0%	34.99%	
	4	-4672.4	15	9374.8	9447.58	0.06	50.42%	1.16%	13.11%	35.31%
Compliance	1	-4679.37	6	9370.73	9399.84	1.00	100%			
	2	-4678.34	9	9374.69	9418.36	0.51	99.47%	0.53%		
	3	-4678.35	12	9380.69	9438.92	0.11	95.03%	0.53%	4.44%	
	4	-4678.07	15	9386.14	9458.92	0.39	18.5%	0%	81.4%	0.11%
Ree	1	-4210.84	6	8433.68	8462.68	1.00	100%			
	2	-4209.16	9	8436.33	8479.82	0.45	99.03%	0.97%		
	3	-4209.16	12	8442.33	8500.31	0.14	95.79%	0%	4.21%	
	4	-4209.17	15	8448.34	8520.82	0.04	0%	0%	83.5%	16.5%
Xee	1	-4180.13	6	8372.26	8401.25	1.00	100%			
	2	-4164.91	9	8347.81	8391.3	0.60	8.09%	91.91%		
	3	-4131.31	12	8286.62	8344.6	0.76	7.66%	91.59%	0.76%	
	4	-4131.31	15	8292.62	8365.1	0.62	8.41%	90.83%	0%	0.76%

Table A.8: Summary table: Univariate airwave oscillometry quadratic models depicted in Figure A.11

Variable	G	loglik	npm	AIC	BIC	entropy	Class 1 (%)	Class 2 (%)	Class 3 (%)	Class 4 (%)
Resistance	1	-4669.32	7	9352.64	9386.61	1.00	100%			
	2	-4661.63	11	9345.27	9398.64	0.22	24.95%	75.05%		
	3	-4661.14	15	9352.29	9425.07	0.46	12.68%	86.89%	0.42%	
	4	-4661.23	19	9360.45	9452.65	0.17	22.94%	0%	74.95%	2.11%
Compliance	1	-4678.35	7	9370.69	9404.66	1.00	100%			
	2	-4672.83	11	9367.65	9421.03	0.89	99.15%	0.85%		
	3	-4672.48	15	9374.96	9447.74	0.55	1.06%	97.36%	1.59%	
	4	-4669.69	19	9377.38	9469.57	0.55	0.85%	0%	97.67%	1.48%
Ree	1	-4209.22	7	8432.45	8466.27	1.00	100%			
	2	-4205.81	11	8433.63	8486.78	0.37	5.72%	94.28%		
	3	-4204.77	15	8439.54	8512.02	0.64	3.67%	96.12%	0.22%	
	4	-4205.37	19	8448.75	8540.55	0.24	5.50%	0%	93.74%	0.76%
Xee	1	-4167.18	7	8348.36	8382.19	1.00	100%			
	2	-4156.70	11	8335.41	8388.56	0.42	15.86	84.14%		
	3	-4105.04	15	8240.08	8312.56	0.67	16.94	82.63%	0.43%	
	4	-4097.53	19	8233.07	8324.88	0.61	2.05%	71.63%	25.89%	0.43%

A.2.4 Univariate LCMM outcomes for MBW and tidal breathing outcomes

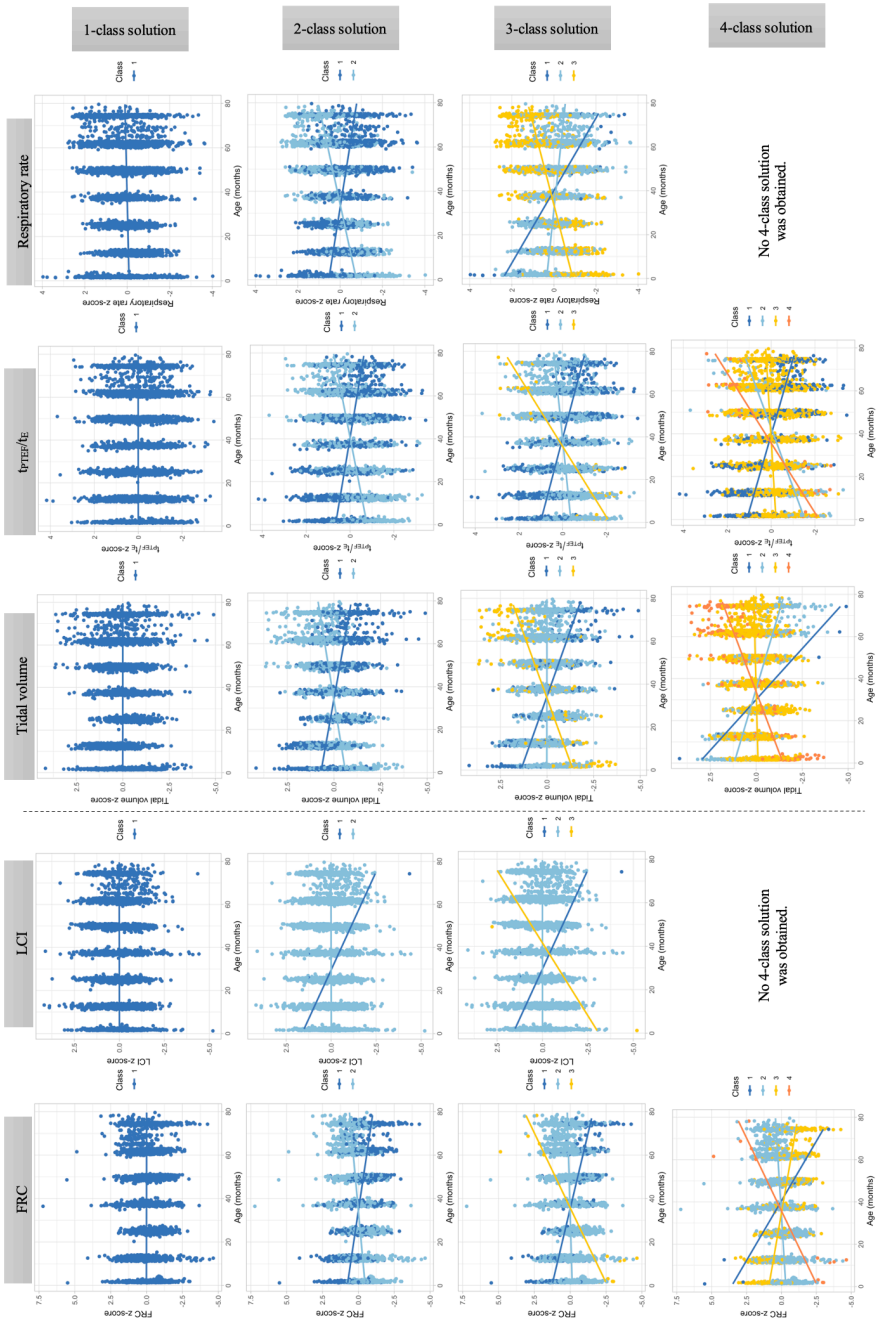


Figure A.12: This figure illustrates the heterogeneous structures within longitudinal univariate MBW (left) and tidal breathing (right) outcomes using linear models. Points show individual scores, while classes were computed through the loess function in R for visualisation purposes.

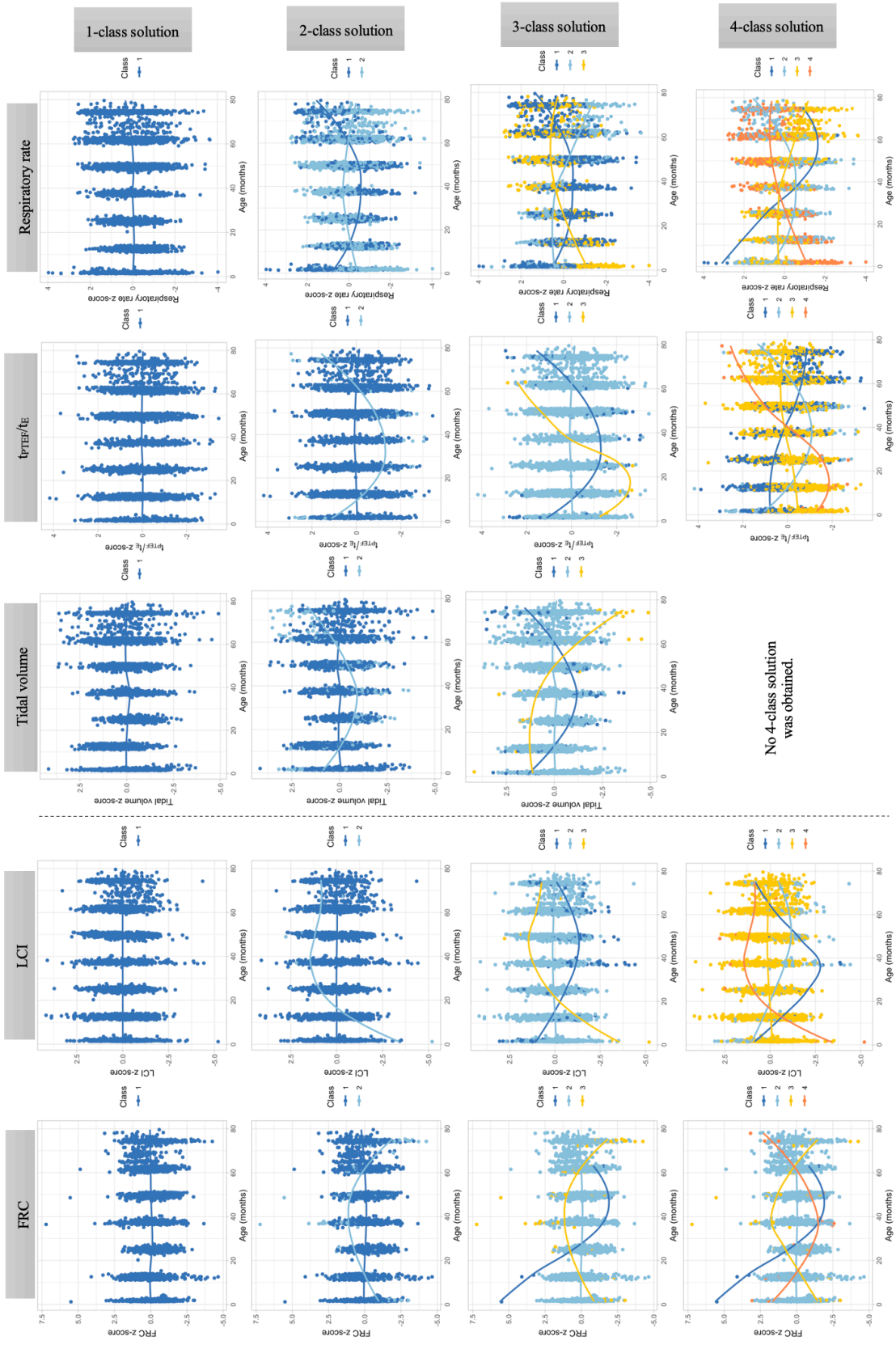


Figure A.13: This figure illustrates the heterogeneous structures within longitudinal univariate MBW (left) and tidal breathing (right) outcomes using quadratic models. Points show individual scores, while classes were computed through the loess function in R for visualisation purposes.

Table A.9: Summary table: Univariate MBW and tidal breathing linear models depicted in Figure A.12

Variable	G	loglik	npm	AIC	BIC	entropy	Class 1 (%)	Class 2 (%)	Class 3 (%)	Class 4 (%)
FRC	1	-5437.32	6	10886.63	10915.83	1.00	100.00	NA	NA	NA
	2	-5412.95	9	10843.89	10887.68	0.23	37.33	62.67	NA	NA
	3	-5407.92	12	10839.85	10898.24	0.60	10.64	88.43	0.94	NA
	4	-5401.96	15	10833.93	10906.91	0.60	0.31	73.62	25.13	0.94
LCI	1	-5437.35	6	10886.70	10915.89	1.00	100.00			
	2	-5436.48	9	10890.95	10934.74	0.51	0.10	99.90		
	3	-5434.48	12	10892.97	10951.36	0.93	0.10	99.69	0.21	
	4	-5434.48	15	10898.96	10971.95	0.93	0.00	99.69	0.10	0.21
Tidal volume	1	-6046.10	6	12104.19	12133.42	1.00	100.00			
	2	-5954.97	9	11927.94	11971.78	0.38	47.20	52.80		
	3	-5936.53	12	11897.05	11955.50	0.53	9.75	80.08	10.17	
	4	-5925.48	15	11880.96	11954.03	0.59	0.31	15.77	75.41	8.51
t_{PTEF}/t_E	1	-6036.15	6	12084.29	12113.52	1.00	100.00			
	2	-5992.50	9	12002.99	12046.83	0.29	53.22	46.78		
	3	-5988.25	12	12000.50	12058.95	0.50	27.59	71.06	1.35	
	4	-5988.18	15	12006.35	12079.42	0.37	22.41	5.19	71.16	1.24
Respiratory rate	1	-6024.41	6	12060.81	12090.04	1.00	100.00			
	2	-5937.09	9	11892.19	11936.03	0.39	53.63	46.37		
	3	-5925.98	12	11875.96	11934.41	0.58	1.56	63.07	35.37	
	4	-5925.98	15	11881.96	11955.02	0.39	1.76	55.19	0.00	43.05

Table A.10: Summary table: Univariate MBW and tidal breathing quadratic models depicted in Figure A.13

Variable	G	loglik	npm	AIC	BIC	entropy	Class 1 (%)	Class 2 (%)	Class 3 (%)	Class 4 (%)
FRC	1	-5408.83	7	10831.66	10865.72	1.00	100.00	NA	NA	NA
	2	-5399.96	11	10821.93	10875.45	0.46	95.62	4.38		
	3	-5386.22	15	10802.44	10875.43	0.71	0.31	96.56	3.13	
	4	-5384.41	19	10806.81	10899.27	0.73	0.31	97.39	1.36	0.94
LCI	1	-5436.43	7	10886.86	10920.93	1.00	100.00			
	2	-5429.24	11	10880.48	10934.01	0.97	99.58	0.42		
	3	-5425.93	15	10881.86	10954.85	0.57	6.15	93.53	0.31	
	4	-5425.02	19	10888.05	10980.50	0.59	0.31	9.38	89.89	0.42
Tidal volume	1	-5932.99	7	11879.98	11914.08	1.00	100.00			
	2	-5925.18	11	11872.35	11925.94	0.30	89.63	10.37		
	3	-5917.85	15	11865.70	11938.77	0.63	3.94	94.92	1.14	
	4	-5917.85	19	11873.70	11966.25	0.39	6.12	92.63	1.24	0.00
t_{PTEF}/t_E	1	-5989.54	7	11993.07	12027.17	1.00	100.00			
	2	-5983.45	11	11988.91	12042.49	0.54	96.68	3.32		
	3	-5981.76	15	11993.52	12066.59	0.69	3.42	96.37	0.21	
	4	-5980.20	19	11998.39	12090.94	0.44	27.49	6.85	64.94	0.73
Respiratory rate	1	-5924.63	7	11863.26	11897.35	1.00	100.00			
	2	-5906.77	11	11835.54	11889.12	0.21	32.47	67.53		
	3	-5903.72	15	11837.43	11910.50	0.24	48.65	21.37	29.98	
	4	-5897.54	19	11833.08	11925.63	0.44	0.62	32.68	37.24	29.46

Bibliography

[1] Li Liu, Francisco Villavicencio, Diana Yeung, Jamie Perin, Gerard Lopez, Kathleen L Strong, and Robert E Black. National, regional, and global causes of mortality in 5–19-year-olds from 2000 to 2019: a systematic analysis. *The Lancet Global Health*, 10(3):e337–e347, 2022.

[2] Heather J Zar and Thomas W Ferkol. The global burden of respiratory diseaseimpact on child health, 2014.

[3] UNICEF et al. United nations childrens fund, committing to child survival: a promise renewed, progress report 2013. erişim tarihi: 25 mart 2016, 2013.

[4] David Rindskopf. The use of latent class analysis in medical diagnosis. In *Recent advances in biostatistics: false discovery rates, survival analysis, and related topics*, pages 257–270. World Scientific, 2011.

[5] Ian D Pavord, Richard Beasley, Alvar Agusti, Gary P Anderson, Elisabeth Bel, Guy Brusselle, Paul Cullinan, Adnan Custovic, Francine M Ducharme, John V Fahy, et al. After asthma: redefining airways diseases. *The Lancet*, 391(10118):350–400, 2018.

[6] Cynthia B Baard, Zoe Franckling-Smith, Jacinta Munro, Lesley Workman, and Heather J Zar. Asthma in south african adolescents: a time trend and risk factor analysis over two decades. *ERJ Open Research*, 7(2), 2021.

[7] R Masekela, CL Gray, Robin J Green, Ahmed I Manjra, FE Kritzinger, M Levin, and H Zar. The increasing burden of asthma in south african children: A call to action. *South African Medical Journal*, 108(7):537–539, 2018.

[8] Anjali Sharma. Respiratory distress. In *Nelson pediatric symptom-based diagnosis*, pages 39–60. Elsevier, 2018.

[9] Sanja Stanojevic, Cole Bowerman, and Paul Robinson. Multiple breath washout: measuring early manifestations of lung pathology. *Breathe*, 17(3), 2021.

[10] Nicole Beydon, Stephanie D Davis, Enrico Lombardi, Julian L Allen, Hubertus GM Arets, Paul Aurora, Hans Bisgaard, G Michael Davis, Francine M Ducharme, Howard Eigen, et al. An

official american thoracic society/european respiratory society statement: pulmonary function testing in preschool children. *American journal of respiratory and critical care medicine*, 175(12):1304–1345, 2007.

[11] Diane M Gray, Dorottya Czovek, Lauren McMillan, Lidija Turkovic, Jacob AM Stadler, Anessa Vanker, Bence L Radics, Zoltán Gingl, Graham L Hall, Peter D Sly, et al. Intra-breath measures of respiratory mechanics in healthy african infants detect risk of respiratory illness in early life. *European Respiratory Journal*, 53(2), 2019.

[12] HJ Zar, W Barnett, L Myer, DJ Stein, and MP Nicol. Investigating the early-life determinants of illness in africa: the drakenstein child health study. *Thorax*, 70(6):592–594, 2015.

[13] Kirsten A Donald, Michelle Hoogenhout, Christopher P du Plooy, Catherine J Wedderburn, Raymond T Nhapi, Whitney Barnett, Nadia Hoffman, Susan Malcolm-Smith, Heather J Zar, and Dan J Stein. Drakenstein child health study (dchs): investigating determinants of early child development and cognition. *BMJ paediatrics open*, 2(1), 2018.

[14] 2017 Socio economic Profile: Drakenstein Municipality (2021). SEP: Socio-economic Profile. `https://www.westerncape.gov.za/provincial-treasury/sites/provincial-treasury.westerncape.gov.za/files/atoms/files/SEP-LG%202021%20-%20Drakenstein%20Municipality.pdf`.

[15] Arthur H Owora and Yijia Zhang. Childhood wheeze trajectory-specific risk factors: A systematic review and meta-analysis. *Pediatric Allergy and Immunology*, 32(1):34–50, 2021.

[16] Rebecca Howard, Magnus Rattray, Mattia Prosperi, and Adnan Custovic. Distinguishing asthma phenotypes using machine learning approaches. *Current allergy and asthma reports*, 15(7):38, 2015.

[17] Andrew J Collaro, Margaret S McElrea, Julie M Marchant, Mark D Chatfield, Peter Sondergeld, Jennifer L Perret, Don Vicendese, Wanaporn Anuntaseree, Shyamali C Dharmage, and Anne B Chang. The effect of early childhood respiratory infections and pneumonia on lifelong lung function: a systematic review. *The Lancet Child & Adolescent Health*, 2023.

[18] James Peter Allinson, Nishi Chaturvedi, Andrew Wong, Imran Shah, Gavin Christopher Donaldson, Jadwiga Anna Wedzicha, and Rebecca Hardy. Early childhood lower respiratory tract infection and premature adult death from respiratory disease in great britain: a national birth cohort study. *The Lancet*, 401(10383):1183–1193, 2023.

[19] Nele Sigurs, Fatma Aljassim, Bengt Kjellman, Paul D Robinson, Fridrik Sigurbergsson, Ragnar Bjarnason, and Per M Gustafsson. Asthma and allergy patterns over 18 years after severe rsv bronchiolitis in the first year of life. *Thorax*, 65(12):1045–1052, 2010.

[20] Adnan Custovic and Sara Fontanella. Evolution of lung function within individuals: clinical insights and data-driven methods, 2023.

[21] Christopher E Troeger, Ibrahim A Khalil, Brigette F Blacker, Molly H Biehl, Samuel B Albertson, Stephanie RM Zimsen, Puja C Rao, Degu Abate, Amha Admasie, Alireza Ahmadi, et al. Quantifying risks and interventions that have affected the burden of lower respiratory infections among children younger than 5 years: an analysis for the global burden of disease study 2017. *The Lancet Infectious Diseases*, 20(1):60–79, 2020.

[22] You Li, Xin Wang, Dianna M Blau, Mauricio T Caballero, Daniel R Feikin, Christopher J Gill, Shabir A Madhi, Saad B Omer, Eric AF Simões, Harry Campbell, et al. Global, regional, and national disease burden estimates of acute lower respiratory infections due to respiratory syncytial virus in children younger than 5 years in 2019: a systematic analysis. *The Lancet*, 399(10340):2047–2064, 2022.

[23] Heather J Zar, Whitney Barnett, Attie Stadler, Sugnet Gardner-Lubbe, Landon Myer, and Mark P Nicol. Aetiology of childhood pneumonia in a well vaccinated south african birth cohort: a nested case-control study of the drakenstein child health study. *The Lancet Respiratory Medicine*, 4(6):463–472, 2016.

[24] Heather J Zar, Polite Nduru, Jacob AM Stadler, Diane Gray, Whitney Barnett, Maia Lesosky, Landon Myer, and Mark P Nicol. Early-life respiratory syncytial virus lower respiratory tract infection in a south african birth cohort: epidemiology and effect on lung health. *The Lancet Global health*, 8(10):e1316–e1325, 2020.

[25] Carlyle McCready, Sadia Haider, Francesca Little, Mark P Nicol, Lesley Workman, Diane M Gray, Raquel Granell, Dan J Stein, Adnan Custovic, and Heather J Zar. Early childhood wheezing phenotypes and determinants in a south african birth cohort: longitudinal analysis of the drakenstein child health study. *The Lancet Child & Adolescent Health*, 7(2):127–135, 2023.

[26] Dan J Stein, N Koen, KA Donald, Colleen M Adnams, S Koopowitz, C Lund, A Marais, B Myers, A Roos, K Sorsdahl, et al. Investigating the psychosocial determinants of child health in africa: The drakenstein child health study. *Journal of neuroscience methods*, 252:27–35, 2015.

[27] WHO ASSIST Working Group. The alcohol, smoking and substance involvement screening test (assist): development, reliability and feasibility. *Addiction*, 97(9):1183–1194, 2002.

[28] Polly FM Robinson, Sara Fontanella, Sachin Ananth, Aldara Martin Alonso, James Cook, Daphne Kaya-de Vries, Luisa Polo Silveira, Lisa Gregory, Clare Lloyd, Louise Fleming, et al. Recurrent severe preschool wheeze: from prespecified diagnostic labels to underlying endotypes. *American journal of respiratory and critical care medicine*, 204(5):523–535, 2021.

[29] BD Spycher, M Silverman, and CE Kuehni. Phenotypes of childhood asthma: are they real? *Clinical & Experimental Allergy*, 40(8):1130–1141, 2010.

[30] Claudia E Kuehni et al. Phenotype specific treatment of obstructive airways disease in infancy and childhood: new recommendations of the swiss paediatric pulmonology group. *Swiss Medical Weekly*, 135(7-8):95–100, 2005.

[31] RM Busquets Monge, A Escribano Montaner, M Fernández Benítez, L García-Marcos, J Garde Garde, M Ibero Iborra, L Pardos Rocamora, J Sánchez Jiménez, E Sánchez Sánchez, J Sanz Ortega, et al. Consensus statement on the management of paediatric asthma. *Allergologia et immunopathologia*, 34(3):88–101, 2006.

[32] Danielle Belgrave and Adnan Custovic. The importance of being earnest in epidemiology. 2016.

[33] Howard Williams and Kenneth N McNicol. Prevalence, natural history, and relationship of wheezy bronchitis and asthma in children. an epidemiological study. *Br Med J*, 4(5679):321–325, 1969.

[34] Fernando D Martinez, Anne L Wright, Lynn M Taussig, Catharine J Holberg, Marilyn Halonen, Wayne J Morgan, and Group Health Medical Associates. Asthma and wheezing in the first six years of life. *New England Journal of Medicine*, 332(3):133–138, 1995.

[35] Christina G Kwong and Leonard B Bacharier. Phenotypes of wheezing and asthma in preschool children. *Current opinion in allergy and clinical immunology*, 19(2):148, 2019.

[36] Abigail Fraser, Corrie Macdonald-Wallis, Kate Tilling, Andy Boyd, Jean Golding, George Davey Smith, John Henderson, John Macleod, Lynn Molloy, Andy Ness, et al. Cohort profile: the avon longitudinal study of parents and children: Alspac mothers cohort. *International journal of epidemiology*, 42(1):97–110, 2013.

[37] Raquel Granell, A John Henderson, and Jonathan A Sterne. Associations of wheezing phenotypes with late asthma outcomes in the avon longitudinal study of parents and children: a population-based birth cohort. *Journal of Allergy and Clinical Immunology*, 138(4):1060–1070, 2016.

[38] Daan Caudri, Olga EM Savenije, Henriette A Smit, Dirkje S Postma, Gerard H Koppelman, Alet H Wijga, Marjan Kerkhof, Ulrike Gehring, Maarten O Hoekstra, Bert Brunekreef, et al. Perinatal risk factors for wheezing phenotypes in the first 8 years of life. *Clinical & Experimental Allergy*, 43(12):1395–1405, 2013.

[39] Olga E Savenije, Raquel Granell, Daan Caudri, Gerard H Koppelman, Henriëtte A Smit, Alet Wijga, Johan C de Jongste, Bert Brunekreef, Jonathan A Sterne, Dirkje S Postma, et al. Comparison of childhood wheezing phenotypes in 2 birth cohorts: Alspac and piama. *Journal of Allergy and Clinical Immunology*, 127(6):1505–1512, 2011.

[40] Souheil Hallit, Benedicte Leynaert, Marie Christine Delmas, Steffi Rocchi, Jacques De Blic, Christophe Marguet, Emeline Scherer, Marie Noelle Dufourg, Corinne Bois, Gabriel Reboux, et al. Wheezing phenotypes and risk factors in early life: The elfe cohort. *PLoS One*, 13(4):e0196711, 2018.

[41] Caroline J Lodge, Sophie Zaloumis, Adrian J Lowe, Lyle C Gurrin, Melanie C Matheson, Christine Axelrad, Catherine M Bennett, David J Hill, Clifford S Hosking, Cecilie Svanes, et al. Early-life risk factors for childhood wheeze phenotypes in a high-risk birth cohort. *The Journal of pediatrics*, 164(2):289–294, 2014.

[42] W Atkinson, J Harris, P Mills, S Moffat, C White, O Lynch, M Jones, P Cullinan, and AJ Newman Taylor. Domestic aeroallergen exposures among infants in an english town. *European Respiratory Journal*, 13(3):583–589, 1999.

[43] RJ Kurukulaaratchy, MH Fenn, LM Waterhouse, SM Matthews, ST Holgate, and SH Arshad. Characterization of wheezing phenotypes in the first 10 years of life. *Clinical & Experimental Allergy*, 33(5):573–578, 2003.

[44] Adnan Custovic, Bridget M Simpson, Clare S Murray, Lesley Lowe, Ashley Woodcock, NAC Manchester Asthma, and Allergy Study Group. The national asthma campaign manchester asthma and allergy study. *Pediatric Allergy and Immunology*, 13:32–37, 2002.

[45] Sheelagh Martindale, Geraldine McNeill, Graham Devereux, Doris Campbell, George Russell, and Anthony Seaton. Antioxidant intake in pregnancy in relation to wheeze and eczema in the first two years of life. *American journal of respiratory and critical care medicine*, 171(2):121–128, 2005.

[46] Global Initiative for Asthma. Global strategy for asthma management and prevention 2023. 2023.

[47] Sadia Haider, Raquel Granell, John Curtin, Sara Fontanella, Alex Cucco, Stephen Turner, Angela Simpson, Graham Roberts, Clare S Murray, John W Holloway, et al. Modeling wheezing spells identifies phenotypes with different outcomes and genetic associates. *American journal of respiratory and critical care medicine*, 205(8):883–893, 2022.

[48] John Henderson, Raquel Granell, Jon Heron, Andrea Sherriff, Angela Simpson, Ashley Woodcock, David P Strachan, Seif O Shaheen, and Jonathan AC Sterne. Associations of wheezing phenotypes in the first 6 years of life with atopy, lung function and airway responsiveness in mid-childhood. *Thorax*, 63(11):974–980, 2008.

[49] Ceyda Oksel, Raquel Granell, Sadia Haider, Sara Fontanella, Angela Simpson, Steve Turner, Graham Devereux, Syed Hasan Arshad, Clare S Murray, Graham Roberts, et al. Distinguishing wheezing phenotypes from infancy to adolescence. a pooled analysis of five birth cohorts. *Annals of the American Thoracic Society*, 16(7):868–876, 2019.

[50] Joanne E Sordillo, Brent A Coull, Sheryl L Rifas-Shiman, Ann Chen Wu, Sharon M Lutz, Marie-France Hivert, Emily Oken, and Diane R Gold. Characterization of longitudinal wheeze phenotypes from infancy to adolescence in project viva, a prebirth cohort study. *Journal of Allergy and Clinical Immunology*, 145(2):716–719, 2020.

[51] Ceyda Oksel, Raquel Granell, Osama Mahmoud, Adnan Custovic, A John Henderson, et al. Causes of variability in latent phenotypes of childhood wheeze. *Journal of Allergy and Clinical Immunology*, 143(5):1783–1790, 2019.

[52] Brian Hallmark, Ganesa Wegienka, Suzanne Havstad, Dean Billheimer, Dennis Ownby, Eneida A Mendonca, Lisa Gress, Debra A Stern, Jocelyn Biagini Myers, Gurjit K Khurana Hershey, et al. Chromosome 17q12-21 variants are associated with multiple wheezing phenotypes in childhood. *American journal of respiratory and critical care medicine*, 203(7):864–870, 2021.

[53] Diane Gray, Lauren Willemse, Ane Visagie, Emilee Smith, Dorottya Czoevek, Peter D Sly, Zoltán Hantos, Graham L Hall, and Heather J Zar. Lung function and exhaled nitric oxide in healthy unsedated a frican infants. *Respirology*, 20(7):1108–1114, 2015.

[54] Cindy T McEvoy, Peter N Le Souef, and Fernando D Martinez. The role of lung function in determining which children develop asthma. *The Journal of Allergy and Clinical Immunology: In Practice*, 11(3):677–683, 2023.

[55] Theresa W Guilbert, Debra A Stern, Wayne J Morgan, Fernando D Martinez, and Anne L Wright. Effect of breastfeeding on lung function in childhood and modulation by maternal asthma and atopy. *American journal of respiratory and critical care medicine*, 176(9):843–848, 2007.

[56] Leslie Elliott, John Henderson, Kate Northstone, Grace Y Chiu, David Dunson, and Stephanie J London. Prospective study of breast-feeding in relation to wheeze, atopy, and bronchial hyperresponsiveness in the avon longitudinal study of parents and children (alspac). *Journal of Allergy and Clinical Immunology*, 122(1):49–54, 2008.

[57] Rosalind J Wright. Perinatal stress and early life programming of lung structure and function. *Biological psychology*, 84(1):46–56, 2010.

[58] Matias Torrent, Jordi Sunyer, Raquel Garcia, Jessica Harris, Maria V Iturriaga, Carme Puig, Oriol Vall, Josep M Anto, Anthony J Newman Taylor, and Paul Cullinan. Early-life allergen exposure and atopy, asthma, and wheeze up to 6 years of age. *American journal of respiratory and critical care medicine*, 176(5):446–453, 2007.

[59] Osama Mahmoud, Raquel Granell, Gabriela P Peralta, Judith Garcia-Aymerich, Deborah Jarvis, John Henderson, and Jonathan Sterne. Early-life and health behaviour influences on lung function in early adulthood. *European Respiratory Journal*, 61(3), 2023.

[60] Fabio Cibella, Andreina Bruno, Giuseppina Cuttitta, Salvatore Bucchieri, Mario Raphael Melis, Stefano De Cantis, Stefania La Grutta, and Giovanni Viegi. An elevated body mass index increases lung volume but reduces airflow in italian schoolchildren. *PLoS One*, 10(5):e0127154, 2015.

[61] Erica S Schultz, Jenny Hallberg, Tom Bellander, Anna Bergström, Matteo Bottai, Flaminia Chiesa, Per M Gustafsson, Olena Gruzieva, Per Thunqvist, Göran Pershagen, et al. Early-life exposure to traffic-related air pollution and lung function in adolescence. *American journal of respiratory and critical care medicine*, 193(2):171–177, 2016.

[62] Vânia Rocha, Sara Soares, Silvia Stringhini, and Sílvia Fraga. Socioeconomic circumstances and respiratory function from childhood to early adulthood: a systematic review and meta-analysis. *BMJ open*, 9(6):e027528, 2019.

[63] Janet Stocks, Alison Hislop, and Samatha Sonnappa. Early lung development: lifelong effect on respiratory health and disease. *The lancet Respiratory medicine*, 1(9):728–742, 2013.

[64] Diane Gray, Lauren Willemse, Ane Visagie, Dorottya Czovek, Polite Nduru, Aneesa Vanker, Dan J Stein, Nastassja Koen, Peter D Sly, Zoltán Hantos, et al. Determinants of early-life lung function in african infants. *Thorax*, 72(5):445–450, 2017.

[65] Diane M Gray, Lidija Turkovic, Lauren Willemse, Ane Visagie, Aneesa Vanker, Dan J Stein, Peter D Sly, Graham L Hall, and Heather J Zar. Lung function in african infants in the drakenstein child health study. impact of lower respiratory tract illness. *American journal of respiratory and critical care medicine*, 195(2):212–220, 2017.

[66] Johnny YC Chan, Debra A Stern, Stefano Guerra, Anne L Wright, Wayne J Morgan, and Fernando D Martinez. Pneumonia in childhood and impaired lung function in adults: a longitudinal study. *Pediatrics*, 135(4):607–616, 2015.

[67] David A McAllister, Li Liu, Ting Shi, Yue Chu, Craig Reed, John Burrows, Davies Adeloye, Igor Rudan, Robert E Black, Harry Campbell, et al. Global, regional, and national estimates of pneumonia morbidity and mortality in children younger than 5 years between 2000 and 2015: a systematic analysis. *The Lancet Global Health*, 7(1):e47–e57, 2019.

[68] Baoting He, Albert M Li, Man Ki Kwok, Shiu Lun Au Yeung, Gabriel M Leung, and C Mary Schooling. Association of growth patterns during infancy and puberty with lung function, wheezing and asthma in adolescents aged 17.5 years: evidence from children of 1997hong kong chinese birth cohort. *International Journal of Epidemiology*, 52(2):440–449, 2023.

[69] Philip H Quanjer, Sanja Stanojevic, Tim J Cole, Xaver Baur, Graham L Hall, Bruce H Culver, Paul L Enright, John L Hankinson, Mary SM Ip, Jinping Zheng, et al. Multi-ethnic reference values for spirometry for the 3–95-yr age range: the global lung function 2012 equations, 2012.

[70] TJ Cole, S Stanojevic, J Stocks, AL Coates, JL Hankinson, and AM Wade. Age-and size-related reference ranges: A case study of spirometry through childhood and adulthood. *Statistics in medicine*, 28(5):880–898, 2009.

[71] Moritz Herle, Nadia Micali, Mohamed Abdulkadir, Ruth Loos, Rachel Bryant-Waugh, Christopher Hübel, Cynthia M Bulik, and Bianca L De Stavola. Identifying typical trajectories in longitudinal data: modelling strategies and interpretations. *European journal of epidemiology*, 35(3):205–222, 2020.

[72] Charles Fletcher and Richard Peto. The natural history of chronic airflow obstruction. *Br Med J*, 1(6077):1645–1648, 1977.

[73] Cristine E Berry, Dean Billheimer, Isaac C Jenkins, Zhenqiang J Lu, Debra A Stern, Lynn B Gerald, Tara F Carr, Stefano Guerra, Wayne J Morgan, Anne L Wright, et al. A distinct low lung function trajectory from childhood to the fourth decade of life. *American journal of respiratory and critical care medicine*, 194(5):607–612, 2016.

[74] Michael J McGeachie, Katherine P Yates, Xiaobo Zhou, Feng Guo, Alice L Sternberg, Mark L Van Natta, Robert A Wise, Stanley J Szefler, Sunita Sharma, Alvin T Kho, et al. Patterns of growth and decline in lung function in persistent childhood asthma. *New England Journal of Medicine*, 374(19):1842–1852, 2016.

[75] Danielle CM Belgrave, Raquel Granell, Steve W Turner, John A Curtin, Iain E Buchan, Peter N Le Souëf, Angela Simpson, A John Henderson, and Adnan Custovic. Lung function trajectories from pre-school age to adulthood and their associations with early life factors: a retrospective analysis of three population-based birth cohort studies. *The Lancet Respiratory Medicine*, 6(7):526–534, 2018.

[76] Joo-Hee Kim, Hun Soo Chang, Seung Woo Shin, Dong Gyu Baek, Ji-Hye Son, Choon-Sik Park, and Jong-Sook Park. Lung function trajectory types in never-smoking adults with asthma: clinical features and inflammatory patterns. *Allergy, asthma & immunology research*, 10(6):614–627, 2018.

[77] Wilfried Karmaus, Nandini Mukherjee, Vimala Devi Janjanam, Su Chen, Hongmei Zhang, Graham Roberts, Ramesh J Kurukulaaratchy, and Hasan Arshad. Distinctive lung function trajectories from age 10 to 26 years in men and women and associated early life risk factors–a birth cohort study. *Respiratory research*, 20:1–13, 2019.

[78] Alex R Horsley, John Belcher, Katie Bayfield, Brooke Bianco, Steve Cunningham, Catherine Fullwood, Andrew Jones, Anna Shawcross, Jaclyn A Smith, Anirban Maitra, et al. Longitudinal assessment of lung clearance index to monitor disease progression in children and adults with cystic fibrosis. *Thorax*, 77(4):357–363, 2022.

[79] Francesca Sanna, Francesca Locatelli, Peter D Sly, Elisha White, David Blake, Jane Heyworth, Graham L Hall, and Rachel E Foong. Characterisation of lung function trajectories and associated early-life predictors in an australian birth cohort study. *ERJ Open Research*, 8(1), 2022.

[80] Raquel Granell, Sadia Haider, Matea Deliu, Anhar Ullah, Osama Mahmoud, Sara Fontanella, Lesley Lowe, Angela Simpson, James William Dodd, Seyed Hasan Arshad, et al. Lung function trajectories from school age to adulthood and their relationship with markers of cardiovascular disease risk. *thorax*, 2024.

[81] Ingrid ML Augustin, Martijn A Spruit, Sarah Houben-Wilke, Frits ME Franssen, Lowie EGW Vanfleteren, Swetlana Gaffron, Daisy JA Janssen, and Emiel FM Wouters. The respiratory physiome: clustering based on a comprehensive lung function assessment in patients with copd. *PLoS One*, 13(9):e0201593, 2018.

[82] Chenjun Ling and Elizabeth C Delmelle. Classifying multidimensional trajectories of neighbourhood change: a self-organizing map and k-means approach. *Annals of GIS*, 22(3):173–186, 2016.

[83] Kimberly J Petersen, Pamela Qualter, and Neil Humphrey. The application of latent class analysis for investigating population child mental health: A systematic review. *Frontiers in psychology*, 10:1214, 2019.

[84] Sarah O'Connor, Claudia Blais, Miceline Mésidor, Denis Talbot, Paul Poirier, and Jacinthe Leclerc. Great diversity in the utilization and reporting of latent growth modeling approaches in type 2 diabetes: A literature review. *Heliyon*, page e10493, 2022.

[85] Nikhil R Garge, Grier P Page, Alan P Sprague, Bernard S Gorman, and David B Allison. Reproducible clusters from microarray research: whither? In *BMC bioinformatics*, volume 6, pages 1–11. BioMed Central, 2005.

[86] Ali Seyed Shirkhorshidi, Saeed Aghabozorgi, and Teh Ying Wah. A comparison study on similarity and dissimilarity measures in clustering continuous data. *PloS one*, 10(12):e0144059, 2015.

[87] Michael R Anderberg. *Cluster analysis for applications: probability and mathematical statistics: a series of monographs and textbooks*, volume 19. Academic press, 2014.

[88] Robert R Sokal and Charles D Miehener. A statistical method for evaluating systematic relationships. *University of Kansas science bulletin*, 38:1409–1438, 1958.

[89] Paul Jaccard. Contribution au problème de l'immigration post-glaciare de la flore alpine. *Bull Soc Vaudoise Sci Nat*, 36:87–130, 1900.

[90] Tom AB Snijders, Maarten Dormaar, Wijbrandt H Van Schuur, Chantal Dijkman-Caes, and Ger Driessen. Distribution of some similarity coefficients for dyadic binary data in the case of associated attributes. *Journal of Classification*, 7(1):5–31, 1990.

[91] Sony Hartono Wijaya, Farit Mochamad Afendi, Irmanida Batubara, Latifah K Darusman, Md Altaf-Ul-Amin, and Shigehiko Kanaya. Finding an appropriate equation to measure similarity between binary vectors: case studies on indonesian and japanese herbal medicines. *BMC bioinformatics*, 17(1):1–19, 2016.

[92] John C Gower. A general coefficient of similarity and some of its properties. *Biometrics*, pages 857–871, 1971.

[93] David Wishart. K-means clustering with outlier detection, mixed variables and missing values. In *Exploratory Data Analysis in Empirical Research*, pages 216–226. Springer, 2003.

[94] Zhexue Huang. Clustering large data sets with mixed numeric and categorical values. In *Proceedings of the 1st pacific-asia conference on knowledge discovery and data mining,(PAKDD)*, pages 21–34. Citeseer, 1997.

[95] János Podani. *Introduction to the exploration of multivariate biological data*. Backhuys Publishers, 2000.

[96] Sandhya Harikumar and PV Surya. K-medoid clustering for heterogeneous datasets. *Procedia Computer Science*, 70:226–237, 2015.

[97] Amir Ahmad and Lipika Dey. A k-mean clustering algorithm for mixed numeric and categorical data. *Data & Knowledge Engineering*, 63(2):503–527, 2007.

[98] Wei Wu, Eugene Bleecker, Wendy Moore, William W Busse, Mario Castro, Kian Fan Chung, William J Calhoun, Serpil Erzurum, Benjamin Gaston, Elliot Israel, et al. Unsupervised phenotyping of severe asthma research program participants using expanded lung data. *Journal of Allergy and Clinical Immunology*, 133(5):1280–1288, 2014.

[99] John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1):100–108, 1979.

[100] Weksi Budiaji and Friedrich Leisch. Simple k-medoids partitioning algorithm for mixed variable data. *Algorithms*, 12(9):177, 2019.

[101] Kaufman L Rousseeuw PJ. Finding groups in data: an introduction to cluster analysis. *Hoboken NJ John Wiley & Sons Inc*, 1990.

[102] Leonard Kaufman and Peter J Rousseeuw. Clustering large applications (program clara). *Finding groups in data: an introduction to cluster analysis*, pages 126–163, 2008.

[103] Raymond T. Ng and Jiawei Han. Clarans: A method for clustering objects for spatial data mining. *IEEE transactions on knowledge and data engineering*, 14(5):1003–1016, 2002.

[104] Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 2009.

[105] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[106] Duc Thanh Anh Luong and Varun Chandola. A k-means approach to clustering disease progressions. In *2017 IEEE International conference on healthcare informatics (ICHI)*, pages 268–274. IEEE, 2017.

[107] T Warren Liao. Clustering of time series dataa survey. *Pattern recognition*, 38(11):1857–1874, 2005.

[108] Paul F Lazarsfeld. The logical and mathematical foundation of latent structure analysis. *Studies in Social Psychology in World War II Vol. IV: Measurement and Prediction*, pages 362–412, 1950.

[109] Leo A Goodman. Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61(2):215–231, 1974.

[110] David J Bartholomew and Martin Knott. Latent variable models and factor analysis, volume 7 of kendalls library of statistics. *Arnold,*, 1999.

[111] Tim J Croudace, Marjo-Riitta Jarvelin, Michael EJ Wadsworth, and Peter B Jones. Developmental typology of trajectories to nighttime bladder control: epidemiologic application of longitudinal latent class analysis. *American Journal of Epidemiology*, 157(9):834–842, 2003.

[112] Linda M Collins and Stephanie T Lanza. *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences*, volume 718. John Wiley & Sons, 2009.

[113] Jose G Dias and Jeroen K Vermunt. A bootstrap-based aggregate classifier for model-based clustering. *Computational Statistics*, 23(4):643–659, 2008.

[114] Baibing Li. A new approach to cluster analysis: the clustering-function-based method. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):457–476, 2006.

[115] Kenneth P Burnham and David R Anderson. Multimodel inference: understanding aic and bic in model selection. *Sociological methods & research*, 33(2):261–304, 2004.

[116] Marie Herr, Jocelyne Just, Lydia Nikasinovic, Christophe Foucault, Anne-Marie Le Marec, Jean-Pierre Giordanella, and Isabelle Momas. Risk factors and characteristics of respiratory and allergic phenotypes in early childhood. *Journal of Allergy and Clinical Immunology*, 130(2):389–396, 2012.

[117] Grazyna Bochenek, Joanna Kuschill-Dziurda, Krystyna Szafraniec, Hanna Plutecka, Andrzej Szczeklik, and Ewa Nizankowska-Mogilnicka. Certain subphenotypes of aspirin-exacerbated respiratory disease distinguished by latent class analysis. *Journal of allergy and clinical immunology*, 133(1):98–103, 2014.

[118] Frances L Garden, Judy M Simpson, Guy B Marks, and Caps Investigators. Atopy phenotypes in the childhood asthma prevention study (caps) cohort and the relationship with allergic disease: clinical mechanisms in allergic disease. *Clinical & Experimental Allergy*, 43(6):633–641, 2013.

[119] Lidia Panico, Beth Stuart, Mel Bartley, and Yvonne Kelly. Asthma trajectories in early childhood: identifying modifiable factors. *PLoS One*, 9(11):e111922, 2014.

[120] John S Uebersax. Probit latent class analysis with dichotomous or ordered category measures: Conditional independence/dependence models. *Applied Psychological Measurement*, 23(4):283–297, 1999.

[121] Yinsheng Qu, Ming Tan, and Michael H Kutner. Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. *Biometrics*, pages 797–810, 1996.

[122] Martin Maechler, Peter Rousseeuw, Anja Struyf, Mia Hubert, and Kurt Hornik. *cluster*. R package version 2.1.1, 2021.

[123] Wendy C Moore, Deborah A Meyers, Sally E Wenzel, W Gerald Teague, Huashi Li, Xingnan Li, Ralph D'Agostino Jr, Mario Castro, Douglas Curran-Everett, Anne M Fitzpatrick, et al. Identification of asthma phenotypes using cluster analysis in the severe asthma research program. *American journal of respiratory and critical care medicine*, 181(4):315–323, 2010.

[124] Anne M Fitzpatrick, W Gerald Teague, Deborah A Meyers, Stephen P Peters, Xingnan Li, Huashi Li, Sally E Wenzel, Shean Aujla, Mario Castro, Leonard B Bacharier, et al. Heterogeneity of severe asthma in childhood: confirmation by cluster analysis of children in the national institutes of health/national heart, lung, and blood institute severe asthma research program. *Journal of allergy and clinical immunology*, 127(2):382–389, 2011.

[125] Mattia CF Prosperi, Umit M Sahiner, Danielle Belgrave, Cansin Sackesen, Iain E Buchan, Angela Simpson, Tolga S Yavuz, Omer Kalayci, and Adnan Custovic. Challenges in identifying asthma subgroups using unsupervised statistical learning techniques. *American journal of respiratory and critical care medicine*, 188(11):1303–1312, 2013.

[126] Drew A Linzer, Jeffrey B Lewis, et al. polca: An r package for polytomous variable latent class analysis. *Journal of statistical software*, 42(10):1–29, 2011.

[127] Ulrike Von Luxburg. Clustering stability: an overview. 2010.

[128] Estibaliz Gómez-de Mariscal, Vanesa Guerrero, Alexandra Sneider, Hasini Jayatilaka, Jude M Phillip, Denis Wirtz, and Arrate Muñoz-Barrutia. Use of the p-values as a size-dependent function to address practical differences when analyzing large datasets. *Scientific reports*, 11(1):20942, 2021.

[129] Cheryl L Grady, Jenny R Rieck, Daniel Nichol, Karen M Rodrigue, and Kristen M Kennedy. Influence of sample size and analytic approach on stability and interpretation of brain-behavior correlations in task-related fmri data. *Human Brain Mapping*, 42(1):204–219, 2021.

[130] Frederick J Rubner, Daniel J Jackson, Michael D Evans, Ronald E Gangnon, Christopher J Tisler, Tressa E Pappas, James E Gern, and Robert F Lemanske Jr. Early life rhinovirus wheezing, allergic sensitization, and asthma risk at adolescence. *Journal of Allergy and Clinical Immunology*, 139(2):501–507, 2017.

[131] Robert F Lemanske Jr, Daniel J Jackson, Ronald E Gangnon, Michael D Evans, Zhanhai Li, Peter A Shult, Carol J Kirk, Erik Reisdorf, Kathy A Roberg, Elizabeth L Anderson, et al. Rhinovirus illnesses during infancy predict subsequent childhood wheezing. *Journal of Allergy and Clinical Immunology*, 116(3):571–577, 2005.

[132] A Vanker, RP Gie, and HJ Zar. The association between environmental tobacco smoke exposure and childhood respiratory disease: a review. *Expert review of respiratory medicine*, 11(8):661–673, 2017.

[133] Sima K Ramratnam, Cynthia M Visness, Katy F Jaffee, Gordon R Bloomberg, Meyer Kattan, Megan T Sandel, Robert A Wood, James E Gern, and Rosalind J Wright. Relationships among maternal stress and depression, type 2 responses, and recurrent wheezing at age 3 years in low-income urban families. *American journal of respiratory and critical care medicine*, 195(5):674–681, 2017.

[134] Rae P MacGinty, Maia Lesosky, Whitney Barnett, Dan J Stein, and Heather J Zar. Associations between maternal mental health and early child wheezing in a south african birth cohort. *Pediatric pulmonology*, 53(6):741–754, 2018.

[135] Symon M Kariuki, Diane M Gray, Charles RJC Newton, Aneesa Vanker, Rae P MacGinty, Nastassja Koen, Whitney Barnett, Lori Chibnik, Karestan C Koenen, Dan J Stein, et al. Association between maternal psychological adversity and lung function in south african infants: a birth cohort study. *Pediatric Pulmonology*, 55(1):236–244, 2020.

[136] David Beran, Heather J Zar, Christophe Perrin, Ana M Menezes, and Peter Burney. Burden of asthma and chronic obstructive pulmonary disease and access to essential medicines in low-income and middle-income countries. *The Lancet Respiratory Medicine*, 3(2):159–170, 2015.

[137] Shaakira Chaya, Rae MacGinty, Carvern Jacobs, Leah Githinji, Sipho Hlengwa, Shannon J Simpson, Heather J Zar, Zoltan Hantos, and Diane M Gray. Normal values of respiratory oscillometry in south african children and adolescents. *ERJ Open Research*, 9(2), 2023.

[138] Paul D Robinson, Philipp Latzin, Sylvia Verbanck, Graham L Hall, Alexander Horsley, Monika Gappa, Cindy Thamrin, Hubertus GM Arets, Paul Aurora, Susanne I Fuchs, et al. Consensus statement for inert gas washout measurement using multiple-and single-breath tests, 2013.

[139] Rachel E Foong, Tim Rosenow, Shannon J Simpson, Benjamin Stoklin, Diane Gray, J Jane Pillow, Graham L Hall, and Kathryn A Ramsey. End-inspiratory molar mass step correction for analysis of infant multiple breath washout tests. *Pediatric Pulmonology*, 52(1):10–13, 2017.

[140] Paul D Robinson, Philipp Latzin, Kathryn A Ramsey, Sanja Stanojevic, Paul Aurora, Stephanie D Davis, Monika Gappa, Graham L Hall, Alex Horsley, Renee Jensen, et al. Preschool multiple-breath washout testing. an official american thoracic society technical statement. *American journal of respiratory and critical care medicine*, 197(5):e1–e19, 2018.

[141] Gregory G King, Jason Bates, Kenneth I Berger, Peter Calverley, Pedro L de Melo, Raffaele L Dellacà, Ramon Farré, Graham L Hall, Iulia Ioan, Charles G Irvin, et al. Technical standards for respiratory oscillometry. *European Respiratory Journal*, 55(2), 2020.

[142] Marije Koopman, Pieter Zanen, Cas LJJ Kruitwagen, Cornelis K van der Ent, and Hubertus GM Arets. Reference values for paediatric pulmonary function testing: The utrecht dataset. *Respiratory medicine*, 105(1):15–23, 2011.

[143] Yosuke Wada, Norihiko Goto, Yoshiaki Kitaguchi, Masanori Yasuo, and Masayuki Hanaoka. Referential equations for pulmonary diffusing capacity using gamlss models derived from japanese individuals with near-normal lung function. *PLoS One*, 17(7):e0271129, 2022.

[144] Timothy J Cole and Pamela J Green. Smoothing reference centile curves: the lms method and penalized likelihood. *Statistics in medicine*, 11(10):1305–1319, 1992.

[145] Kenneth J Rothman, Sander Greenland, Timothy L Lash, et al. *Modern epidemiology*, volume 3. Wolters Kluwer Health/Lippincott Williams & Wilkins Philadelphia, 2008.

[146] Robert A Rigby and D Mikis Stasinopoulos. Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(3):507–554, 2005.

[147] Mikis D Stasinopoulos, Robert A Rigby, Gillian Z Heller, Vlasios Voudouris, and Fernanda De Bastiani. *Flexible regression and smoothing: using GAMLSS in R*. CRC Press, 2017.

[148] Robert A Rigby and DM Stasinopoulos. A semi-parametric additive model for variance heterogeneity. *Statistics and Computing*, 6:57–65, 1996.

[149] Peter J Green and Bernard W Silverman. *Nonparametric regression and generalized linear models: a roughness penalty approach*. Crc Press, 1993.

[150] Anton Frederik Thielmann, René-Marcel Kruse, Thomas Kneib, and Benjamin Säfken. Neural additive models for location scale and shape: A framework for interpretable neural regression beyond the mean. In *International Conference on Artificial Intelligence and Statistics*, pages 1783–1791. PMLR, 2024.

[151] George EP Box and David R Cox. An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2):211–243, 1964.

[152] Anthony Biglan, Dennis Ary, and Alexander C Wagenaar. The value of interrupted time-series experiments for community intervention research. *Prevention Science*, 1:31–49, 2000.

[153] Anita K Wagner, Stephen B Soumerai, Fang Zhang, and Dennis Ross-Degnan. Segmented regression analysis of interrupted time series studies in medication use research. *Journal of clinical pharmacy and therapeutics*, 27(4):299–309, 2002.

[154] Cécile Proust-Lima, Viviane Philipps, and Benoit Liquet. Estimation of extended mixed models using latent classes and latent processes: the r package lcmm. *arXiv preprint arXiv:1503.00890*, 2015.

[155] Gavin van der Nest, Valéria Lima Passos, Math JJM Candel, and Gerard JP van Breukelen. An overview of mixture modelling for latent evolutions in longitudinal data: Modelling approaches, fit statistics and software. *Advances in Life Course Research*, 43:100323, 2020.

[156] Sandra H van Oostrom, Peter M Engelfriet, WM Monique Verschuren, Maarten Schipper, Inge M Wouters, Marike Boezen, Henriëtte A Smit, Huib AM Kerstjens, and H Susan J Picavet. Aging-related trajectories of lung function in the general populationthe doetinchem cohort study. *PLoS One*, 13(5):e0197250, 2018.

[157] Carlyle McCready, Heather J Zar, Shaakira Chaya, Carvern Jacobs, Lesley Workman, Zoltan Hantos, Graham L Hall, Peter D Sly, Mark P Nicol, Dan J Stein, et al. Determinants of lung function development from birth to age 5 years: an interrupted time series analysis of a south african birth cohort. *The Lancet Child & Adolescent Health*, 8(6):400–412, 2024.

[158] Alvar Agusti and Rosa Faner. Lung function trajectories in health and disease. *The Lancet Respiratory Medicine*, 7(4):358–364, 2019.

[159] Brian K Jordan and Cindy T McEvoy. Trajectories of lung function in infants and children: setting a course for lifelong lung health. *Pediatrics*, 146(4), 2020.

[160] Lieke Voncken, Casper J Albers, and Marieke E Timmerman. Model selection in continuous test norming with gamlss. *Assessment*, 26(7):1329–1346, 2019.

[161] Yoav Benjamini, Dan Drai, Greg Elmer, Neri Kafkafi, and Ilan Golani. Controlling the false discovery rate in behavior genetics research. *Behavioural brain research*, 125(1-2):279–284, 2001.

[162] Diane M Gray, Catherine J Wedderburn, Rae P MacGinty, Lauren McMillan, Carvern Jacobs, Jacob AM Stadler, Graham L Hall, and Heather J Zar. Impact of hiv and antiretroviral drug exposure on lung growth and function over 2 years in an african birth cohort. *Aids*, 34(4):549–558, 2020.

[163] Katherine L O'Brien, Henry C Baggett, W Abdullah Brooks, Daniel R Feikin, Laura L Hammitt, Melissa M Higdon, Stephen RC Howie, Maria Deloria Knoll, Karen L Kotloff, Orin S Levine, et al. Causes of severe pneumonia requiring hospital admission in children without hiv infection from africa and asia: the perch multi-country case-control study. *The Lancet*, 394(10200):757–779, 2019.

[164] Giorgia Bussu, Emily JH Jones, Tony Charman, Mark H Johnson, and Jan K Buitelaar. Latent trajectories of adaptive behaviour in infants at high and low familial risk for autism spectrum disorder. *Molecular Autism*, 10:1–12, 2019.

[165] Osvaldo Espin-Garcia, Lizbeth Naranjo, and Ruth Fuentes-García. A latent class linear mixed model for monotonic continuous processes measured with error. *Statistical Methods in Medical Research*, 33(3):449–464, 2024.

[166] John A Reynolds, Jennifer Prattley, Nophar Geifman, Mark Lunt, Caroline Gordon, Ian N Bruce, and MASTERPLANS Consortium. Distinct patterns of disease activity over time in patients with active sle revealed using latent class trajectory models. *Arthritis research & therapy*, 23:1–12, 2021.

[167] G van der Nest. *Hidden depths: robustness of modelling approaches for uncovering latent classes in longitudinal data*. Doctoral thesis, Maastricht University, 2022.

[168] James TD Gibbons, Andrew C Wilson, and Shannon J Simpson. Predicting lung health trajectories for survivors of preterm birth. *Frontiers in Pediatrics*, 8:318, 2020.

[169] Cassidy Du Berry, Christopher Nesci, Jeanie LY Cheong, Tara FitzGerald, Rheanna Mainzer, Sarath Ranganathan, Lex W Doyle, Elianne JLE Vrijlandt, and Liam Welsh. Long-term expiratory airflow of infants born moderate-late preterm: a systematic review and meta-analysis. *EClinicalMedicine*, 52, 2022.

[170] Mònica Guxens, Agnes MM Sonnenschein-van der Voort, Henning Tiemeier, Albert Hofman, Jordi Sunyer, Johan C de Jongste, Vincent WV Jaddoe, and Liesbeth Duijts. Parental psychological distress during pregnancy and wheezing in preschool children: the generation r study. *Journal of Allergy and Clinical Immunology*, 133(1):59–67, 2014.

[171] AL Phelan, MR DiBenedetto, IM Paul, J Zhu, and KH Kjerulff. Psychosocial stress during first pregnancy predicts infant health outcomes in the first postnatal year. *Maternal and child health journal*, 19:2587–2597, 2015.

[172] D Mikis Stasinopoulos and Robert A Rigby. Generalized additive models for location scale and shape (gamlss) in r. *Journal of Statistical Software*, 23:1–46, 2008.

[173] Bobby L Jones and Daniel S Nagin. Advances in group-based trajectory modeling and an sas procedure for estimating them. *Sociological methods & research*, 35(4):542–571, 2007.

[174] Daniel S Nagin, Bobby L Jones, Valeria Lima Passos, and Richard E Tremblay. Group-based multi-trajectory modeling. *Statistical methods in medical research*, 27(7):2015–2023, 2018.

[175] Sara Dolnicar, Bettina Grün, and Friedrich Leisch. Increasing sample size compensates for data problems in segmentation studies. *Journal of Business Research*, 69(2):992–999, 2016.

[176] MJ Greenacre. Theory and applications of correspondence analysis, 1984.

[177] Michael Greenacre. *Correspondence analysis in practice*. chapman and hall/crc, 2017.