Pediatric Pulmonology

WILEY

ORIGINAL ARTICLE OPEN ACCESS

Clinical Profile Identification of Indigenous Infants With Bronchiolitis Through Using Unsupervised Feature Extraction and Clustering

Hongqi Niu[1] | Gabrielle Britt McCallum[2] | Anne Bernadette Chang[2,3] | Khalid Khan[4] | Sami Azam[1]

[1]Faculty of Science and Technology, Charles Darwin University, Darwin, Northern Territory, Australia | [2]Child and Maternal Health Division, Menzies School of Health Research, Charles Darwin University, Darwin, Northern Territory, Australia | [3]Department of Respiratory and Sleep Medicine, Queensland Children's Hospital and Australian Centre for Health Services Innovation (AusHSI), Queensland University of Technology, Brisbane, Queensland, Australia | [4]Faculty of Arts and Society, Charles Darwin University, Darwin, Northern Territory, Australia

Correspondence: Hongqi Niu (hongqi.niu@cdu.edu.au)

Received: 24 April 2025 | **Revised:** 21 November 2025 | **Accepted:** 19 December 2025

Funding: Research Training Programme (RTP) Fees Offset and Stipend Scholarship from the Australian Commonwealth Government; NHMRC Leadership (L3), Grant/Award Number: 2025379

Keywords: Bronchiectasis | Bronchiolitis | Dimensionality reduction | Phenotyping | Small datasets | Unsupervised feature extraction

ABSTRACT

Objective: Infants hospitalized with bronchiolitis may experience persistent symptoms linked to future chronic lung diseases like bronchiectasis. Identifying phenotypes during hospitalization could guide targeted interventions. As traditional clustering requires large datasets, this study explores whether Unsupervised Feature Extraction Algorithms (UFEAs) and clustering can identify high-risk profiles in a small dataset of Indigenous infants.

Methods: We included 128 Indigenous infants hospitalized with bronchiolitis at the Royal Darwin Hospital, Northern Territory, Australia. Eight UEFAs were applied to reduce the dimensionality of 22 variables across 2–17 dimensions. A support vector machine classifier assessed the effectiveness of each UFEA in classifying bronchiectasis. Kernel Principal Component Analysis with nine dimensions performed best, and these dimensions were used for clustering.

Results: Six clinical profiles were identified. Profile C, the highest-risk group with the most infants with bronchiectasis (45%), preterm birth (95%), low birth weight (86%), weight-for-length z-score < -2 (62%), household smoke exposure (90%), and antibiotics prescribed before hospitalization (100%). Profile D, the second-highest risk, had bronchiectasis (30%), the highest wet/productive cough (45%), crackles/crepitations (36%), and wheeze (18%). Profile F infants included bronchiectasis (22%), oxygen supplementation (91%), and lobar collapse/consolidation on chest X-rays (65%). Profile A included bronchiectasis (5%) and household smoke exposure (30%), and Profile E showed bronchiectasis (9%) and household smoke exposure (36%). Profile B, the lowest-risk group, with no bronchiectasis (0%), preterm birth (15%), low birth weight (10%), and any bacteria (5%).

Conclusion: Using UFEAs and clustering, we reduced dataset dimensionality, effectively identifying six unique, clinically significant risk profiles in Indigenous infants.

1 | Introduction

Worldwide, bronchiolitis continues to be a leading cause of hospitalization for infants, with more than 3.6 million episodes annually [1]. Bronchiolitis is a heterogeneous, multi-dimensional disorder characterised by varying clinical phenotypes and potential differences in pathophysiology, risk factors, and outcomes [2–5]. While typically self-limiting, some infants

have persistent respiratory symptoms beyond-hospitalization that are associated with ongoing respiratory morbidity and future bronchiectasis [6, 7]. Childhood bronchiectasis is now recognized globally as a disease of importance, with the prevalence estimated to range from 0.2 to 735 per 100,000 children [8, 9]. A particularly high disease burden and severity of bronchiectasis however is observed among Indigenous infants within high-income countries [10], such as Australian First Nations, as well as in children from low-middle income countries [9].

Identifying phenotypic subgroups in infants with respiratory disorders such as asthma [11] and bronchiectasis [8] would thus allow for tailoring treatment to improve clinical outcomes in both the short- and longer-term. Methods using clustering, omics approaches, and machine learning are valuable tools to categorize children based on risk profiles that may enable earlier intervention facilitation to prevent chronic respiratory diseases [12, 13]. However, typically, such analyses require large datasets, especially when there are many dimensions (i.e. a high number of variables) to the disease and/or outcomes. High-Dimensional Small-Sample Size (HDSSS) [14, 15] datasets present a major challenge to these techniques as analysis is more complex [12]. One way to overcome this challenge is through using 'Dimension Reduction' (DR) techniques that aim to simplify complex datasets and improve data quality by minimizing the number of input variables [16]. DR techniques are generally divided into two types: (a) feature selection [17], e.g., Multiple Correspondence Analysis (MCA) which selects the most relevant features while discarding the less significant ones; and (b) feature extraction [16], which creates a lower-dimensional representation of the data while retaining essential information. Unsupervised Feature Extraction Algorithms (UFEAs) are particularly useful in HDSSS [18, 19] as unsupervised methods can identify hidden patterns in data without reliance on labelled datasets. This makes them well-suited for real-life datasets exhibiting noise, complexity, and sparsity.

Previous research involving infants with bronchiolitis using Latent Class Analysis (LCA) [13, 20–22] and MCA identified distinct profiles associated with increased risk of future asthma [13, 20], bronchiectasis [12] and wheezing [13, 20, 22]. Yet, these approaches only involved feature selection, which may lead to the loss of valuable information by excluding unselected features. For example, in our previous MCA-LCA study [12], only 12 variables were retained from a possible 22 from the dataset, potentially overlooking important insights factors associated with bronchiectasis.

Our current study shifts focus from feature selection to feature extraction to account for these previous limitations. Utilizing UFEAs, we analysed data previously collected from Australian Indigenous infants hospitalized with bronchiolitis [12, 23–25]. This dataset, derived from an HDSSS, uses all 22 variables while reducing dimensionality to facilitate meaningful clustering and risk profile identification. Our hypothesis is that applying UFEAs for dimensionality reduction without excluding variables makes it possible to identify high-risk profiles that differ from those identified in our previous study. This approach preserves all the variables with reduced dimensions, enabling us to explore whether the extracted profiles provide novel insights into risk factors associated with longer-term outcomes of bronchiolitis among Indigenous infants. Our aim was to determine whether applying UEFAs and clustering on a small multi-dimensional dataset of Indigenous infants hospitalized with bronchiolitis can identify high-risk profiles.

2 | Methods

2.1 | Summary of Original Studies

We used de-identified data from our previous three studies involving Indigenous infants aged < 2 years hospitalized with bronchiolitis, recruited between June 2008 and September 2013 from the Royal Darwin Hospital, Darwin, Australia [23–25]. Two were randomized controlled trials (RCTs) evaluating whether one or three doses of weekly azithromycin, compared to placebo, improved clinical outcomes for infants hospitalized with bronchiolitis [23, 25]. The third was a cohort study that evaluated the validity and reliability of a bronchiolitis scoring system for infants admitted to hospital with bronchiolitis [24]. All infants had clinical, viral, and bacterial data collected [23–25]. Data relating to bronchiectasis were obtained from a separate study at the same tertiary hospital as described previously [7]. Each of these original studies received approval from the local Human Research Ethics Committee (HREC 07/60, HREC-2010-1324), and written informed consent was obtained from the primary caregivers. As this current study represents a reanalysis of existing data, no additional ethical approval was required.

In our previous MCA-LCA [12] study, 164 infants and 22 variables were included. In this study, we replaced variables 'caregiver reported cough (last 7 days)' and 'caregiver reported breathing difficulty (last 7 days)' with variables 'presence of cough' and 'chest auscultation abnormalities' respectively identified by a medical professional, as these latter variables were considered more objective and clinically relevant. We then removed participants with any missing data, leaving 128 infants for analysis. Further methods are described in the original studies [23–25] and summarised in our previous study [12].

2.2 | Statistical Analysis

Demographics were described using median and interquartile range (IQR: 25-75th percentile) for continuous variables, while categorical variables were presented as frequency and percentages. Weight-for-length z-score categories were established using Zanthro software in STATA v14.0, and data analysis was conducted in Python. We then undertook the 'Dimensionality Selection Process' followed by clustering analysis to achieve our study aims. We also compared the findings of this study with the findings from our previous study MCA-LCA [12].

Virus and bacteria data were obtained from nasopharyngeal swabs (NPS) collected at enrolment. NPS were placed into skim milk tryptone glucose glycerol broth (STGGB), stored at -80°C. NPS bacterial pathogens were cultured at our institution [26] and respiratory viruses using PCR undertaken at the Queensland Paediatric Infectious Diseases laboratory in Brisbane, as previously done [23–25]. For this study, we combined any bacteria detected as a single variable and only the two most frequently detected viruses (RSV and HRV) were analysed (so as to allow inclusion of more data).

TABLE 1 | Nine dimensions performance comparison.

Dim = 9	Accuracy	F1 Score	Precision	Sensitivity	False positive rate	False negative rate	False discovery rate	Negative predictive value
PCA	0.846	0.625	0.63	0.625	0.097	0.375	0.375	0.903
MDS	0.918	0.787	0.83	**0.763**	0.042	**0.238**	0.163	**0.941**
KPCA	**0.949**	**0.857**	**1**	**0.750**	**0.000**	**0.250**	**0**	**0.939**
Isomap	0.795	0.429	0.5	0.375	0.097	0.625	0.375	0.848
LLE	0.795	nan	nan	0	0.000	1	0	0.795
LE	0.769	0.182	0.33	0.125	0.065	0.875	0.25	0.806
FastICA	0.846	0.625	0.625	0.625	0.097	0.375	0.375	0.903
Autoencoder	0.797	nan	nan	0.175	0.042	0.825	0.163	0.821

TP = True Positive (when the model correctly Identified as having HD).
TN = True Negative (when the model correctly identified the opposite class, such as patients truly having no heart issues).
FP = False Positive (when the model incorrectly identified HD patients i.e., identifying non-HD patients as HD patients).
FN = False Negative (when the model incorrectly identified the opposite class, such as HD patients as normal patients).
Accuracy = (TP + TN)/(TP + TN + FP + FN).
Precision = (TP)/(TP + FP).
Sensitivity = (TP)/(TP + FN).
F1 Score = 2(Precision X Sensitivity)/(Precision + Sensitivity).
False Positive Rate = FP/(FP + TN).
False Negative Rate = FN/(TP + FN).
False Discovery Rate = FP/(FN + TP).
Negative predictive value = TN/(TN + FN).

2.3 | Dimensionality Selection Process

We applied eight unsupervised feature extraction algorithms (UEFAs)—Principal Component Analysis (PCA), Classical Multidimensional Scaling (MDS), Kernel Principal Component Analysis (KPCA), Isomap, Locally Linear Embedding (LLE), Laplacian Eigenmaps (LE), Fast Independent Component Analysis (FastICA), and Autoencoder—to reduce the dimensionality of data while retaining the information of all 22 variables. Each UFEA was run iteratively to reduce the dataset into dimensions ranging from 2 to 17, representing up to 80% of the maximum possible dimensions.

To evaluate the effectiveness of each UFEA, we used a support vector machine (SVM) classifier to assess the ability of the reduced dimensions to classify bronchiectasis. This approach ensured that the reduced dimensions retained critical information relevant to the classification task.

The SVM classification accuracy for bronchiectasis (with bronchiectasis labels) was used solely to evaluate the performance of dimensionality reduction algorithms, helping to identify the best-performing method. This evaluation does not influence the clustering results. Among the eight algorithms, KPCA, with nine dimensions, achieved the best performance for classifying bronchiectasis. Table 1 illustrates the results with standard evaluation metrics.

These nine dimensions extracted by KPCA, which provided an optimal representation of the dataset while preserving the relationships among the original 22 variables, were then utilized for clustering analysis.

2.4 | Clustering Analysis

After obtaining nine dimensions from KPCA, we applied four clustering methods—K-means, DBSCAN, OPTICS, and LCA—to group the data and the performance evaluated by using three methods: Davies-Bouldin Index (DBI), Calinski-Harabasz Index (CHI), and Silhouette Coefficient Score (SCS) [27]. Table 2 illustrates that K-means achieved the highest silhouette coefficient score and Calinski-Harabasz Index compared to DBSCAN and OPTICS. It showed only a slight difference in the DBI score compared to OPTICS, still indicating the best-defined clusters overall.

TABLE 2 | Performance comparison between clustering methods.

Dimensions	Methods	DBI	CHI	SCS
Reduced 9 as per KPCA	*K*-Means	**1.73**	**16.38**	**0.15**
	Optics	1.94	5.74	−0.05
	DBScan	1.78	6.09	−0.03
No Reduction	LCA	2.59	6.83	0.049

Abbreviations: DBI, Davies–Bouldin Index; CHI, Calinski–Harabasz Index; SCS, Silhouette Coefficient Score.

3 | Results

Table 3 depicts the demographic, medical, and clinical characteristics of our study cohort. 128 Indigenous infants were involved, and 22 variables were selected in this study. The results of the dimensionality reduction experiments are illustrated in Table 1, where among the UEFAs, KPCA with nine dimensions attained the highest results of 94.9% test accuracy for the SVM model. Table 2 shows the clustering performances. The *K*-Means clustering attained the highest results as per CHI (16.38) and SCS (0.15) and DBI of 1.73, only 0.05 difference than DB Scan, makes the K-Means best clustering model for our study and K-Means give six clusters that are considered as six clinical profiles for proving our initial hypothesis.

TABLE 3 | Baseline demographic, medical and clinical characteristics.

Variables	Total $n = 128$ (%)
Median age (months)	3.5 (IQR 3–5)
Aged < 12 months	109 (85)
Boys	82 (64)
Preterm birth (< 37-weeks)	33 (26)
Low birth weight (< 2.5 kg)	31 (24)
Weight-for-length z-score[a]	
> 2	4 (3)
+ 1 to +2	10 (8)
−1 to +1	72 (56)
−2 to −1	21 (16)
< −2	21 (16)
Remote[b]	110 (86)
Currently breastfed	114 (89)
Mother smoked during pregnancy	71 (55)
Exposed to household smoke	79 (62)
Previous respiratory hospitalisation	24 (19)
Number required supplemental oxygen	78 (61)
Antibiotics prescribed prior to hospital	113 (88)
Any co-morbidity[c]	64 (50)
Lobar collapse/consolidation on chest x-ray	26 (20)
Presence of Cough	
Child did not cough	102 (80)
Dry cough	7 (5)
Wet/Productive cough	19 (15)
Chest auscultation	
Any Chest Recession	2 (2)
Crackles/Crepitations	17 (13)
Normal	99 (77)
Wheeze	10 (8)
Accessory muscle use[d] (24)	
+	41 (32)
++	50 (39)
+++	30 (23)
None	7 (5)
Length of stay (hours)	
< 48	40 (31)
48–72	27 (21)
72–96	36 (28)
> 96	25 (20)
Respiratory Syncytial Virus	56 (44)
Human Rhinovirus	36 (28)
Any bacteria[e]	92 (72)
Bronchiectasis on HRCT	27 (21)

Abbreviations: HRCT, high-resolution computed tomography; IQR, interquartile range.
[a]Weight-for-length *z*-score categories are presented as standard deviation;
[b]Remote: e.g. > 100 km from a tertiary hospital;
[c]Any co-morbidity included presence of otitis media or any skin infection;
[d]Accessory muscle use: None (no chest in-drawing); + (presence of mild intercostal in-drawing); ++ (moderate amount of intercostal in-drawing); +++ (moderate or marked intercostal in-drawing with presence of head bobbing or tracheal tug);
[e]Any bacteria detected on nasopharyngeal swab include: Streptococcus pneumoniae, Haemophilus influenza, Moraxella catarrhalis, Staphylococcus aureus. Bronchiectasis on HRCT included 3 categories in the original dataset: 0 = no bronchiectasis; 1 = bronchiectasis confirmed, and 2 = HRCT undertaken but not bronchiectasis identified. In this study, we combined categories 0 and 2 to 0 (no bronchiectasis).

Among the six profiles identified in this current study, two (Profiles C and D) were associated with bronchiectasis that have equal or more than 30% of infants with confirmed bronchiectasis to identify the factors responsible. The remaining four profiles included bronchiectasis in 22%, 9%, 5% and 0% from Profiles F, E, A and B respectively. The overall results for each profile are described in the Table 4. Table 5 shows the differences in profiles from this current study compared to our previous MCA-LCA study [12].

Profile C included 16% of the overall cohort, characterized by the highest rate of bronchiectasis (45%) and many known risk factors such as preterm birth, low birth weight, exposure to household smoke, weight-for-length z-score, and antibiotics prescribed prior to hospital. Importantly, Profile C was consistent with Profile AMCA-LCA from our previous MCA-LCA study, sharing many of the same risk factors [12] with the exception of additional risk factors such as exposure to household smoke and antibiotics prescribed before hospital stay.

Profile D included 26% of the overall cohort, with 30% having confirmed bronchiectasis, wet cough, crackles, wheezing, and previous respiratory hospitalization. Compared with our previous MCA-LCA study, additional factors included the presence of cough and abnormal chest auscultation.

Profile F group included 18% of the overall cohort, with 22% of the cohort with bronchiectasis, which was charactered by a significant need for oxygen supplementation. Lobar collapse or consolidation on chest x-rays was also more frequent, Antibiotics prescribed prior to hospital, and second high exposure to household smoke.

Profile-E group included 9% of the overall cohort, 9% of the cohort with bronchiectasis, outstanding characteristic in this group was high rate of Antibiotics prescribed prior to hospital, Any co-morbidity and Respiratory Syncytial Virus. Although RSV and any co-morbidity were confirmed to be associated with severe symptom of bronchiolitis, this group might protect by the following factors: none of the group are preterm births (0% vs. 0–95%) or low birth weight (0% vs. 0–86%); the highest frequency of normal weight-for-length z-scores (82% vs. 10–64%); second low Exposed to household smoke (36% vs.

TABLE 4 | Summary of the nine dimensions and six profiles.

	Profile A $n = 20$ (16%)	Profile B $n = 20$ (16%)	Profile C $n = 21$ (16%)	Profile D $n = 33$ (26%)	Profile E $n = 11$ (9%)	Profile F $n = 23$ (18%)	Total $n = 128$ (100%)
Age (< 12 months)	109 (100)	109 (90)	109 (86)	109 (64)	109 (91)	109 (96)	109 (85)
BoysB Boys	82 (85)	82 (55)	82 (43)	82 (73)	82 (36)	82 (74)	82 (64)
Preterm (< 37-weeks)	33 (0)	33 (15)	33 (95)	33 (30)	33 (0)	33 (0)	33 (26)
Low birth weight (< 2.5 kg)	31 (0)	31 (10)	31 (86)	31 (27)	31 (0)	31 (9)	31 (24)
Weight-for-length Z-score[a]							
> 2	4 (0)	4 (0)	4 (0)	4 (3)	4 (9)	4 (9)	4 (3)
+1 to +2	10 (10)	10 (15)	10 (0)	10 (12)	10 (0)	10 (4)	10 (8)
−1 to +1	72 (80)	72 (60)	72 (10)	72 (64)	72 (82)	72 (52)	72 (56)
−2 to −1	21 (5)	21 (20)	21 (29)	21 (12)	21 (9)	21 (22)	21 (16)
≤ 2	21 (5)	21 (5)	21 (62)	21 (9)	21 (0)	21 (13)	21 (16)
Remote[b]	110 (70)	110 (100)	110 (100)	110 (67)	110 (100)	110 (96)	110 (86)
Currently breastfed	114 (100)	114 (90)	114 (95)	114 (67)	114 (100)	114 (100)	114 (89)
Mother smoked during pregnancy	71 (45)	71 (85)	71 (43)	71 (48)	71 (18)	71 (78)	71 (55)
Exposed to household smoke	79 (30)	79 (60)	79 (90)	79 (55)	79 (36)	79 (87)	79 (62)
Previous respiratory hospitalisation	24 (5)	24 (15)	24 (24)	24 (33)	24 (0)	24 (17)	24 (19)
Number required supplemental oxygen	78 (45)	78 (70)	78 (67)	78 (45)	78 (45)	78 (91)	78 (61)
Antibiotics prescribed prior to hospital	113 (80)	113 (95)	113 (100)	113 (70)	113 (100)	113 (100)	113 (88)
Any co-morbidity[c]	64 (10)	64 (50)	64 (67)	64 (42)	64 (82)	64 (65)	64 (50)
Lobar collapse/consolidation on x-ray	26 (0)	26 (15)	26 (29)	26 (6)	26 (0)	26 (65)	26 (20)
Presence of Cough							
Child did not cough	102 (95)	102 (100)	102 (81)	102 (45)	102 (100)	102 (87)	102 (80)
Dry cough	7 (0)	7 (0)	7 (10)	7 (9)	7 (0)	7 (9)	7 (5)
Wet/productive cough	19 (5)	19 (0)	19 (10)	19 (45)	19 (0)	19 (4)	19 (15)
Chest auscultation							
Any chest recession	2 (0)	2 (0)	2 (0)	2 (6)	2 (0)	2 (0)	2 (2)
Crackles/crepitations	17 (0)	17 (15)	17 (0)	17 (36)	17 (0)	17 (9)	17 (13)
Wheeze	10 (0)	10 (5)	10 (10)	10 (18)	10 (0)	10 (4)	10 (8)

(Continues)

TABLE 4 | (Continued)

	Profile A $n = 20$ (16%)	Profile B $n = 20$ (16%)	Profile C $n = 21$ (16%)	Profile D $n = 33$ (26%)	Profile E $n = 11$ (9%)	Profile F $n = 23$ (18%)	Total $n = 128$ (100%)
Normal	99 (100)	99 (80)	99 (90)	99 (39)	99 (100)	99 (87)	99 (77)
Accessory muscle use[d]							
+	41 (30)	41 (55)	41 (24)	41 (39)	41 (27)	41 (13)	41 (32)
++	50 (30)	50 (25)	50 (38)	50 (45)	50 (45)	50 (48)	50 (39)
+++	30 (35)	30 (20)	30 (24)	30 (9)	30 (27)	30 (35)	30 (23)
None	7 (5)	7 (0)	7 (14)	7 (6)	7 (0)	7 (4)	7 (5)
Length of stay (hours)							
< 48	40 (10)	40 (60)	40 (33)	40 (39)	40 (18)	40 (17)	40 (31)
48–72	27 (30)	27 (15)	27 (19)	27 (9)	27 (45)	27 (26)	27 (21)
72–96	36 (45)	36 (5)	36 (24)	36 (36)	36 (0)	36 (39)	36 (28)
> 96	25 (15)	25 (20)	25 (24)	25 (15)	25 (36)	25 (17)	25 (20)
Respiratory Syncytial Virus	56 (40)	56 (25)	56 (48)	56 (42)	56 (91)	56 (39)	56 (44)
Human Rhinovirus	36 (50)	36 (20)	36 (24)	36 (36)	36 (27)	36 (9)	36 (28)
Any bacteria[e]	92 (90)	92 (5)	92 (76)	92 (85)	92 (82)	92 (87)	92 (72)
Bronchiectasis on HRCT[f]	27 (5)	27 (0)	27 (45)	27 (30)	27 (9)	27 (22)	27 (21)

Abbreviations: HRCT, high-resolution computed tomography; IQR, interquartile range.
[a]Weight-for-length *z*-score categories are presented as standard deviation;
[b]Remote: e.g. > 100 km from a tertiary hospital;
[c]Any co-morbidity included presence of otitis media or any skin infection;
[d]Accessory muscle use: None (no chest in-drawing); + (presence of mild intercostal in-drawing); ++ (moderate amount of intercostal in-drawing); +++ (moderate or marked intercostal in-drawing with presence of head bobbing or tracheal tug);
[e]Any bacteria detected on nasopharyngeal swab include: Streptococcus pneumoniae, Haemophilus influenza, Moraxella catarrhalis, Staphylococcus aureus.
[f]Bronchiectasis on HRCT has 3 categories in the original dataset, 0, 1 and 2 which 2 means take the CT but didn't not developed Bronchiectasis, based on the result, we set the value 2 to 0, then Bronchiectasis has 2 groups:0 and 1.

TABLE 5 | Comparison with the key differences among profiling with MCA-LCA study (12).

Current Profiles and characteristics ($n = 128$)			Previous MCA-LCA Profiles and characteristics ($n = 164$)		
Profiles $n = 128$(%)	**Key characteristics (current profile vs. other profiles)**	**Bronchiectasis on HRCT $n = 27$ (%)**	**Profiles $n = 164$ (%)**	**Key characteristics (current profile vs. other profiles)**	**Bronchiectasis on HRCT* $n = 34$ (%)**
Profile C $n = 21$(16%)	• Preterm births (95% vs. 0%–30%) • Birth weight (86% vs. 0%–27%) • Weight-for-length *z*-score[a] below −2 (62% vs. 0%–13%) • Household smoke exposure was high (90% vs. 30%–87%) • Antibiotics prescribed prior to hospital (100% vs. 70%–100%)	$n = 10$ (45%)	Profile $C_{MCA\text{-}LCA}$ $n = 11$ (7%)	• bacterial detected[b] (93.1% vs. 56.7–72%) • Any co-morbidity[c] (75.4% vs 33.9–68.3%) • Supplemental oxygen required (100% vs. 20.3% ~ 100%) • Accessory muscle use[d] (++: 84.7% vs. 0–51.4%).	$n = 11$ (100%)
Profile D $n = 33$ (26%)	• Wet or productive cough (45% vs. 0%–10%) • Mild crackles (36% vs. 0%–15%) • Wheezing (18% vs. 0%–15%) • Previous respiratory hospitalizations (33% vs. 0%–24%)	$n = 10$ (30%)	Profile $A_{MCA\text{-}LCA}$ $n = 39$ (23.8%)	• Preterm birth (90.7% vs. 0–12.6%) • Birth weight (89.2% vs. 0–15.0%) • Weight-for-length *z*-score (−2 to −1: 58.8% vs. 0–6.8%) • Previous respiratory hospitalisation (39.6% vs. 11.1–19.4%)	$n = 14$ (35.4%)
Profile F $n = 23$ (18%)	• Oxygen supplementation (91% vs. 45–70%) • Lobar collapse or consolidation on chest x-rays (65% vs. 0%–29%) • Antibiotics prescribed prior to hospital (100% vs. 70%–100%) • Exposure to household smoke (87%)	$n = 5$ (22%)	Profile $E_{MCA\text{-}LCA}$ $n = 53$ (32.2%)	• Respiratory Syncytial Virus detected (44.0%% vs. 32.2–49.4%) • Bacteria detective (72.0% vs. 56.7–93.1%)	$n = 7$ (13.8%)
Profile E $n = 11$ (9%)	• Antibiotics prescribed prior to hospital (100% vs. 70%–100%) • Any co-morbidity (82% vs. 10%–67%) • Respiratory Syncytial Virus (91% vs. 25%–48%)	$n = 1$ (9%)	Profile $D_{MCA\text{-}LCA}$ $n = 19$ (11.6%)	• Human Rhinovirus detected (49.4% vs 23.2–32.5) • Weight-for-length *z*-score (> 2: 26.2% vs. 0%) (+ 1 to +2: 31.7% vs. 0% ~ 9.4%) (< 2: 36% vs. 8.2% ~ 23.9%)	$n = 2$ (7.0%)
Profile A $n = 20$ (16%)	• Human Rhinovirus (50% vs. 9%–36%)	$n = 1$ (5%)	Profile $B_{MCA\text{-}LCA}$ $n = 42$ (25.3%)	• Respiratory Syncytial Virus (48.6% vs. 32.2–44%)	$n = 0$ (0.0%)

(Continues)

TABLE 5 | (Continued)

Current Profiles and characteristics ($n = 128$)			Previous MCA-LCA Profiles and characteristics ($n = 164$)		
Profiles $n = 128$(%)	**Key characteristics (current profile vs. other profiles)**	**Bronchiectasis on HRCT $n = 27$ (%)**	**Profiles $n = 164$ (%)**	**Key characteristics (current profile vs. other profiles)**	**Bronchiectasis on HRCT* $n = 34$ (%)**
Profile B $n = 20$ (16%)	• Any bacteria (90% vs. 5%–87%) Maternal smoking during pregnancy (85% vs. 18%–48%)	$n = 0$ (0%)		• Accessory muscle use (+++: 45.5% vs. 13.4–26.2%) • Supplemental oxygen required (100% vs. 20.3–100%) • Length of stay (hours) (48.1% for 60–96 h; 34.4% for > 96 h)	

'vs' means to compare with the other profiles. For example, in Profile $C_{MCA\text{-}LCA}$, High bacterial detection (93.1% vs. 56.7–72%) means 93.1% bacterial detection in Profile $C_{MCA\text{-}LCA}$, and in the other profiles it ranges from 56.7% to 72%.

Abbreviations:

*HRCT: high-resolution computed tomography.

[a]Weight-for-length *z*-score categories are presented as standard deviation;

[b]Any bacteria include *Streptococcus pneumoniae, Haemophilus influenza, Moraxella catarrhalis, Staphylococcus aureus*;

[c]Any co-morbidity includes: any otitis media or any skin infection;

[d]Accessory muscle use: None (no chest in-drawing); + (presence of mild intercostal in-drawing); ++ (moderate amount of intercostal in-drawing); +++ (moderate or marked intercostal in-drawing with presence of head bobbing or tracheal tug).

30–90%) and Previous respiratory hospitalisation (0% vs. 5–33%).

Profile-A group included 16% of the cohort, 5% of the cohort with bronchiectasis, which was charactered by human Rhinovirus, and Any bacteria. This group might protect by the following factors: none of the group are preterm births (0% vs. 0–95%) or low birth weight (0% vs. 0–86%); second high frequency of normal weight-for-length *z*-scores (80% vs. 10–64%); less exposed to household smoke (30% vs. 36–90%); lowest number of co-morbidities (10% vs. 42–82%); Lobar collapse/consolidation on x-ray (0% vs. 0–65%); Child did not cough (95% vs. 45–100%) and Normal Chest auscultation (100% vs. 39–100%).

Profile-B group included 16% of the cohort, with 0% of the cohort with bronchiectasis, which was characterised by the high maternal smoking rate during pregnancy. Compared with the other profiles, infants in this group had fewer preterm births (15% vs. 0%–95%) or low birth weight (10% vs. 0%–86%); Child did not cough (100% vs. 45%–100%); Length of stay (hours) < 48 (60% vs. 10%–33%); less Any bacteria (5% vs. 79%–90%); less Respiratory Syncytial Virus (25% vs. 39%–91%); second less Human Rhinovirus (20% vs. 9%–50%).

3.1 | Comparison with the Previous MCA-LCA [12] Study

From our current and previous MCA-LCA [12] studies, four clinical profiles were associated with bronchiectasis. In the previous MCA-LCA [12] study, Profile $C_{MCA\text{-}LCA}$ [12], included 100% of the cohort with bronchiectasis, characterized by presence of high bacterial rates (93.1%), presence of co-morbidities (75.4%), requirement for supplemental oxygen (100%), and accessory muscle use (84.7%). Similarly, Profile $A_{MCA\text{-}LCA}$ included 35.4% with bronchiectasis, with high rates of preterm birth (90.7%), low birth weight (89.2%), poor weight-for-length z-score (58.8%) and the highest previous respiratory hospitalization (39.6%).

In contrast to our current study, Profile C included 45% with bronchiectasis, was similar to Profile $A_{MCA\text{-}LCA}$ [12] of the MCA-LCA study [12]. This profile was characterized by the highest frequency of preterm births (95%), low birth weight (86%), poor weight-for-length z-scores (62%), and two new key characteristics, i.e., household smoke exposure (90%) and antibiotics prescribed before hospitalization (100%). Profile D, a new profile found in our current study, comprises 30% of the cohort with bronchiectasis and is characterized by the highest prevalence of wet or productive cough (45%), the presence of crackles (36%), the highest wheezing rate (18%), and the highest rate of previous respiratory hospitalizations (33%). Figure 1 shows the comparison with the previous MCA-LCA study.

4 | Discussion

We have demonstrated the utility of UFEAs in reducing high-dimensional data in a cohort of Indigenous infants hospitalized with bronchiolitis. By applying *K*-means clustering to the reduced dimensions, we identified six clinical profiles, two of which (Profiles C and D) contained the highest proportion of

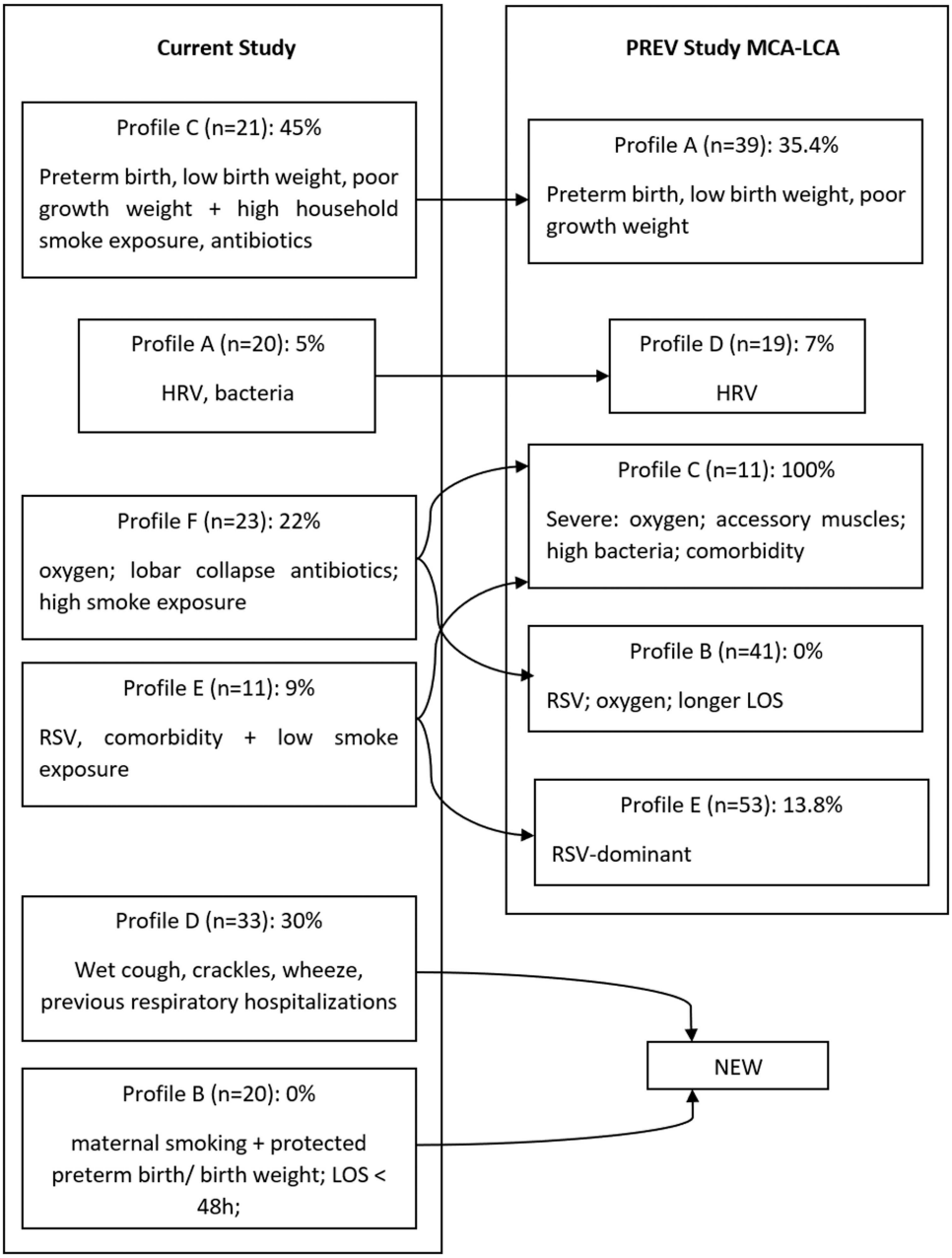

FIGURE 1 | Comparison between the current study and the previous MCA-LCA study. The percentages shown represent the proportion of participants with bronchiectasis on HRCT within each group.

infants with bronchiectasis, offering important insights into associated risk factors for developing bronchiectasis. Compared to traditional methods used for clustering (i.e., LCA), the six profiles in this study provided further depth of data.

Our study is the first to use method UFEAs to examine if additional variables can be included in our clinical dataset (a small dataset), thereby offering a broader understanding of bronchiectasis-associated risks. Uniquely, compared to traditional methods used for clustering, i.e., LCA, except for Profile C, other profiles we identified in this study differed from the clinical profiles (shown in Table 5).

Our study found that Profile C was characterized with the highest rate of bronchiectasis (45%) between the six profiles. It included factors such as preterm birth, low birth weight, poor growth (indicated by a weight-for-length z-score <-2), and a high rate of antibiotics prescribed prior to hospital. Environmental exposures, notably household smoke, were also prominent, suggesting a multifactorial contribution to bronchiectasis risk. This profile highlights the combined influence of early-life vulnerabilities and environmental conditions on long-term respiratory health. Profile C aligns partially with Profile $A_{MCA\text{-}LCA}$ from the MCA-LCA [12] study, as both share risk factors like preterm birth, low birth weight, and poor growth. However, this study identified additional factors, such as household smoke exposure and early-life antibiotic use which were not included in our previous MCA-LCA [12] analysis. These differences underscore the advantage of using UFEAs, which avoid excluding variables based solely on their perceived relevance in feature selection.

In contrast, Profile D included 30% of the cohort with and was defined by symptoms such as a wet or productive cough,

crackles, wheezing, and previous respiratory hospitalizations. Profile D also provides additional insights, incorporating variables such as the presence of a cough and chest auscultation findings, which were not included in the MCA-LCA [12] study. These variables offer further insights for understanding respiratory morbidity and potential development of bronchiectasis.

Importantly, the primary difference between this current study and our previous MCA-LCA [12] study lies in the dimensionality and the number of variables included. In our previous MCA-LCA study [12], the feature selection technique MCA was employed to identify the optimal variables. The MCA-LCA [12] analysis excluded twelve variables that are known risk factors for bronchiectasis, such as exposure to parental or household smoke. In this study, UFEAs allowed all 22 variables to be retained, enabling a more comprehensive analysis.

This study highlights two key advantages of using UFEAs. First, by retaining all variables, UFEAs ensured that no potentially valuable information was lost, allowing for a more detailed exploration of risk factors. Second, this approach facilitated the identification of new variables that were not part of the previous analyses, providing a broader understanding of bronchiectasis-associated risks. Although our study is novel, there are also limitations. A larger dataset could enhance clinical profiling, improve generalizability, and yield more robust, distinguished findings applicable across broader populations.

To the best of our knowledge, this study is the first to utilize feature extraction techniques in analysing a cohort of Indigenous infants hospitalized with bronchiolitis, a group at increased risk of bronchiectasis. Profiles C and D, which included the highest proportion of bronchiectasis cases, reveal crucial insights into the factors contributing to long-term respiratory outcomes. Profile C emphasizes the combined impact of early-life factors, such as preterm birth and low birth weight, and environmental exposures like household smoke, while Profile D highlights the importance of ongoing respiratory symptoms and previous hospitalizations. The findings demonstrate the potential of UFEAs, combined with clustering methods, to identify clinically relevant risk profiles in small, multidimensional datasets. This methodology retains critical variables and enhances the understanding of complex conditions like bronchiolitis and their progression to bronchiectasis. Although UFEA algorithms have advantages in analysing small datasets [28], having more data typically leads to more reliable and generalizable results.

Despite the unique methods described, its clinical utility may be viewed from studies that utilise data-driven methods rather than hypothesis driven. These include using the profiles for more targeted therapy and surveillance e.g., children with high-risk profiles (e.g., Profile C) showed strong associations with bronchiectasis and modifiable risk factors like household smoke exposure could be targeted for more intense follow-up. From an equity lens, our method is important as traditional clustering methods often require large datasets, limiting their applicability in Indigenous or underserved populations. Future research among a larger cohort including a validation cohort could utilize emerging UFEAs to refine phenotyping further and inform targeted interventions to mitigate the burden of bronchiectasis.

Furthermore, while current analyses focus on clinical and phenotypic features, genetic data may be helpful to elucidate geno-phenotyping as evidence from Indigenous populations in North America has shown that integrating genotypes can reveal common immunodeficiencies and ciliary dysfunction underlying respiratory phenotypes [29]. However, our well-established Indigenous Reference Group have recurrently stated they do not conduct genetic studies. Indigenous groups around the globe vary considerably but research must always be respectful and only undertaken with Indigenous-led support and guidance [30].

In summary, we have presented a data-driven method that enables utilisation of small datasets. Using this method, we found 6 profiles; of which 2 had high bronchiectasis prevalence. Whether these children should be targeted with more intensive treatment and/or follow-up require further data analysis.

Author Contributions

Hongqi Niu: conceptualization, writing – original draft, writing – review and editing. **Gabrielle Britt McCallum:** writing – review and editing. **Anne Bernadette Chang:** writing – review and editing. **Khalid Khan:** review and editing. **Sami Azam:** supervision, writing – review and editing. All authors reviewed and approved the final manuscript.

Acknowledgments

We are grateful for the children and families who participated in the original studies. Hongqi Niu is supported by a Research Training Programme (RTP) Fees Offset and Stipend Scholarship from the Australian Commonwealth Government; and Anne Bernadette Chang is funded by a NHMRC Leadership (L3) fellowship (grant 2025379). No other funding was provided for this analysis.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The datasets presented in this article are not available as per our institutions' policies involving Australian First Nations children and in accordance with national guidelines. We are unable to share individual participant data as specific consent for this was not obtained.

References

1. H. Nair, E. A. Simões, I. Rudan, et al., "Global and Regional Burden of Hospital Admissions for Severe Acute Lower Respiratory Infections in Young Children in 2010: A Systematic Analysis," *Lancet* 381, no. 9875 (2013): 1380–1390, https://doi.org/10.1016/S0140-6736(12)61901-1.

2. S. Cunningham, H. Nair, and H. Campbell, "Deciphering Clinical Phenotypes in Acute Viral Lower Respiratory Tract Infection: Bronchiolitis is Not an Island," *BMJ Publishing Group Ltd* 71 (2016): 679–680.

3. B. L. K. Chawes, P. Poorisrisak, S. L. Johnston, and H. Bisgaard, "Neonatal Bronchial Hyperresponsiveness Precedes Acute Severe Viral Bronchiolitis in Infants," *Journal of Allergy and Clinical Immunology* 130, no. 2 (2012): 354–361, https://doi.org/10.1016/j.jaci.2012.04.045.

4. R. C. Welliver, "Review of Epidemiology and Clinical Risk Factors for Severe Respiratory Syncytial Virus (RSV) Infection," *Journal of Pediatrics* 143, no. 5 Suppl (2003): S112–S117, https://doi.org/10.1067/s0022-3476(03)00508-0.

5. E. J. Bailey, C. Maclennan, P. S. Morris, et al., "Risks of Severity and Readmission of Indigenous and Non-Indigenous Children Hospitalised for Bronchiolitis," *Journal of Paediatrics and Child Health* 45, no. 10 (2009): 593–597, https://doi.org/10.1111/j.1440-1754.2009.01571.x.

6. G. B. McCallum, E. J. Plumb, P. S. Morris, and A. B. Chang, "Antibiotics for Persistent Cough or Wheeze Following Acute Bronchiolitis in Children," *Cochrane Database of Systematic Reviews* 8, no. 8 (2017): CD009834.

7. G. B. McCallum, M. D. Chatfield, P. S. Morris, and A. B. Chang, "Risk Factors for Adverse Outcomes of Indigenous Infants Hospitalized with Bronchiolitis," *Pediatric Pulmonology* 51, no. 6 (2016): 613–623, https://doi.org/10.1002/ppul.23342.

8. A. B. Chang, A. Bush, and K. Grimwood, "Bronchiectasis in Children: Diagnosis and Treatment," *Lancet* 392, no. 10150 (2018): 866–879, https://doi.org/10.1016/S0140-6736(18)31554-X.

9. G. B. McCallum and M. J. Binks, "The Epidemiology of Chronic Suppurative Lung Disease and Bronchiectasis in Children and Adolescents," *Frontiers in Pediatrics* 5 (2017): 27.

10. P. C. Valery, P. J. Torzillo, K. Mulholland, N. C. Boyce, D. M. Purdie, and A. B. Chang, "Hospital-Based Case-Control Study of Bronchiectasis in Indigenous Children in Central Australia," *Pediatric Infectious Disease Journal* 23, no. 10 (2004): 902–908.

11. I. D. Pavord, R. Beasley, A. Agusti, et al., "After Asthma: Redefining Airways Diseases," *Lancet* 391, no. 10118 (2018): 350–400, https://doi.org/10.1016/S0140-6736(17)30879-6.

12. H. Niu, A. B. Chang, V. M. Oguoma, Z. Wang, and G. B. McCallum, "Latent Class Analysis to Identify Clinical Profiles Among Indigenous Infants With Bronchiolitis," *Pediatric Pulmonology* 55, no. 11 (2020): 3096–3103.

13. O. Dumas, J. M. Mansbach, T. Jartti, et al., "A Clustering Approach to Identify Severe Bronchiolitis Profiles in Children," *Thorax* 71, no. 8 (2016): 712–718, https://doi.org/10.1136/thoraxjnl-2016-208535.

14. Y. Saeys, I. Inza, and P. Larrañaga, "A Review of Feature Selection Techniques in Bioinformatics," *Bioinformatics* 23, no. 19 (2007): 2507–2517.

15. J. Friedman, T. Hastie, and R. Tibshirani The elements of statistical learning: Springer series in statistics New York 2001.

16. C. M. Bishop and N. M. Nasrabadi, *Pattern Recognition and Machine Learning*, Vol. 4 (Springer, 2006).

17. I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," *Journal of machine learning research 2003* 3, no. Mar: 1157–1182.

18. F. Anowar, S. Sadaoui, and B. Selim, "Conceptual and Empirical Comparison of Dimensionality Reduction Algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, T-SNE)," *Computer Science Review* 40 (2021): 100378.

19. Y. Jun Yan, Z. Benyu Zhang, L. Ning Liu, et al., "Effective and Efficient Dimensionality Reduction for Large-Scale and Streaming Data Preprocessing," *IEEE transactions on Knowledge and Data Engineering* 18, no. 3 (2006): 320–333.

20. O. Dumas, K. Hasegawa, J. M. Mansbach, A. F. Sullivan, P. A. Piedra, and C. A. Camargo, "Severe Bronchiolitis Profiles and Risk of Recurrent Wheeze by Age 3 Years," *Journal of Allergy and Clinical Immunology* 143, no. 4 (2019): 1371–1379, https://doi.org/10.1016/j.jaci.2018.08.043.

21. D. A. Linzer and J. B. Lewis, "poLCA: An R Package for Polytomous Variable Latent Class Analysis," *Journal of Statistical Software* 42, no. 10 (2011): 1–29.

22. L. Petrarca, R. Nenna, G. Di Mattia, et al., "Bronchiolitis Phenotypes Identified by Latent Class Analysis May Influence the Occurrence of Respiratory Sequelae," *Pediatric Pulmonology* 57, no. 3 (2022): 616–622.

23. G. B. McCallum, P. S. Morris, M. D. Chatfield, et al., "A Single Dose of Azithromycin Does Not Improve Clinical Outcomes of Children Hospitalised With Bronchiolitis: A Randomised, Placebo-Controlled Trial," *PLoS One* 8, no. 9 (2013): e74316.

24. G. B. McCallum, P. S. Morris, C. C. Wilson, et al., "Severity Scoring Systems: are They Internally Valid, Reliable and Predictive of Oxygen Use in Children With Acute Bronchiolitis?," *Pediatric Pulmonology* 48, no. 8 (2013): 797–803.

25. G. B. McCallum, P. S. Morris, K. Grimwood, et al., "Three-Weekly Doses of Azithromycin for Indigenous Infants Hospitalized With Bronchiolitis: A Multicentre, Randomized, Placebo-Controlled Trial," *Frontiers in Pediatrics* 3 (2015): 32, https://doi.org/10.3389/fped.2015.00032.

26. K. M. Hare, K. Grimwood, A. J. Leach, et al., "Respiratory Bacterial Pathogens in the Nasopharynx and Lower Airways of Australian Indigenous Children With Bronchiectasis," *Journal of Pediatrics* 157, no. 6 (2010): 1001–1005, https://doi.org/10.1016/j.jpeds.2010.06.002.

27. A. Karim, S. Azam, B. Shanmugam, and K. Kannoorpatti, "An Unsupervised Approach for Content-Based Clustering of Emails into Spam and Ham Through Multiangular Feature Formulation," *IEEE Access* 9 (2021): 135186–135209.

28. H. Niu, G. B. McCallum, A. B. Chang, K. Khan, and S. Azam, "Exploring Unsupervised Feature Extraction Algorithms: Tackling High Dimensionality in Small Datasets," *Scientific Reports* 15, no. 1 (2025): 21973.

29. L. Vicuña, "Genetic Associations with Disease in Populations With Indigenous American Ancestries," *Genetics and Molecular Biology* 47, no. Suppl 1 (2024): e20230024.

30. A. B. Chang, T. Kovesi, G. Redding, et al., "Improving the Respiratory Health of Indigenous Peoples Globally the Need, Why and How," *Lancet Respiratory Medicine* 12, no. 7 (2024): 556–574, https://doi.org/10.1016/S2213-2600(24)00008-0.