Spatial Machine-Learning Prediction of Lung Function and Body Mass Index in Australian Cystic Fibrosis Patients

Simran Chawla[1] Ahmadreza Pourghaderi[2] Arul Earnest[3]

[1, 2, 3]*School of Public Health, Monash University, 553 St Kilda Road, Melbourne, Victoria 3004, Australia*

Correspondence to: Arul Earnest arul.earnest@monash.edu

1. Introduction

Cystic fibrosis (CF) is a debilitating genetic disorder caused by mutations in the CFTR gene, leading to chronic airway infection, and progressive lung damage (1,2). In Australia, more than 3,700 individuals are registered with our Australian Cystic Fibrosis Data Registry (ACFDR), and despite advances in antimicrobial therapy, airway-clearance techniques, nutrition, and CFTR modulators, substantial variation in health outcomes remains (3). Lung function, measured by percent-predicted FEV_1 ($ppFEV_1$%), and nutritional status, captured through BMI z-scores in children, continue to be one of the strongest determinants of morbidity and survival.

Emerging evidence suggests that geographic and contextual factors, such as socioeconomic disadvantage, air quality, greenness, and remoteness, may contribute to spatial inequities in CF outcomes. Traditional modelling approaches also assume independence between observations, which is frequently violated in spatially clustered health data, limiting both accuracy and interpretability (4).

To address these gaps, this study develops a spatial machine-learning framework integrating national ACFDR data with high-resolution environmental and socioeconomic datasets. By predicting $ppFEV_1$% and BMI z-scores using spatially explicit ML models, the work demonstrates how modern AI techniques can enhance precision public health and support more equitable CF service planning across Australia.

2. Methods

2.1 Study Design

We conducted a retrospective spatial ML analysis using features collected longitudinally to predict 2024 outcomes. Separate models were built for:
Adults (≥18 years) – predicting $ppFEV_1$%
Children (5–17 years) – predicting BMI z-score

2.2 Data Sources

Registry data from the ACFDR were linked to environmental and socioeconomic variables including:

- SEIFA IRSD (socioeconomic disadvantage)
- ABS remoteness (RA 2021)
- Walkability (National Liveability Study)
- Greenness (NDVI)
- $PM_{2.5}$ (1 km) satellite-derived annual means
- NO_2 (TROPOMI) tropospheric column densities

All spatial layers were harmonised to the Australian Albers projection (EPSG:3577). Postcode centroids were used to extract environmental exposures. Missing data were imputed within cross-validation folds to improve precision and reduce bias.

2.3 Machine-Learning Framework & Validation

Models were developed using Random Forests (500 trees) implemented in the mlr3 package in R software. Key features:

- Non-linear modelling capability
- Robustness to multicollinearity
- Permutation importance for interpretability
- Predictors included demographic, clinical, infection, genotype, pancreatic function, and contextual variables.

To ensure valid out-of-sample evaluation, we employed 50 km spatial-block cross-validation. Performance metrics included RMSE, MAE, and R^2. The trained 2023 models were then applied to 2024 hold-out data for temporal external validation. Sensitivity analyses tested alternative block sizes and compared Random Forests with XGBoost.

3. Results

3.1 Descriptive

From 1,854 adults and 1,163 children seen in 2023, after excluding incomplete or invalid geographic records, analytic cohorts included:
Adults: 1,845 (train 2023), 1,727 (predict 2024)
Children: 1,048 (train 2023), 1,054 (predict 2024)

3.2 Model Performance

Final model performance was reasonably good and summarised in table 1.

Table 1. Model performance summary (2023–2024)

Cohort	2023 Spatial CV (Internal)	2024 Hold-out Validation	Δ (2024 – 2023)
Adults ($ppFEV_1$ %)	RMSE 19.67 ± 0.91 MAE 15.98 ± 1.00 R^2 0.22 ± 0.05	RMSE 11.70 MAE 9.38 R^2 0.82	↓ RMSE –7.97 ↓ MAE –6.60 ↑ R^2 +0.60
Children (BMI z-score)	RMSE 0.59 ± 0.03 MAE 0.46 ± 0.02 R^2 0.61 ± 0.02	RMSE 0.50 MAE 0.38 R^2 0.77	↓ RMSE –0.09 ↓ MAE –0.08 ↑ R^2 +0.17

3.3 Feature Importance

For adults, lung-function predictions were driven by a mix of clinical and contextual features:

- Age at diagnosis & at measurement
- Pseudomonas aeruginosa infection
- BMI
- Socioeconomic disadvantage (IRSD)
- Environmental Greenness (NDVI)
- Walkability
- Air pollutants ($PM_{2.5}$, NO_2)

For children, BMI z-scores were overwhelmingly determined by intrinsic growth variables (weight, age, height), with minimal contribution from contextual factors.

3.4 Spatial Insights

Spatial distribution of CF patients in our study is shown in figure 1 below. Significant regional variation was observed, with lower lung function more common in Queensland and New South Wales.

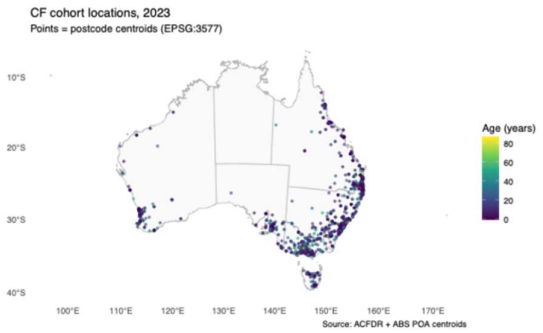

Figure 1. Distribution of CF patients by postcode centroids in Australia

Lower lung function also concentrated in regions with socioeconomic disadvantage and higher NO_2 and $PM_{2.5}$ observed in major urban corridors.

4. Discussion

This study demonstrates the feasibility and value of integrating spatial context into CF outcome prediction at a national scale in Australia. The adult lung-function model captured rich interactions between individual health status, chronic infection, socioeconomic disadvantage, environmental exposures, and regional characteristics. Compared to traditional regression approaches, the spatial ML framework accounted for geographic clustering and dependency, enabling more realistic prediction and reducing overoptimistic accuracy from non-spatial CV.

In contrast, the child BMI model was dominated by growth-related variables, suggesting current area-level predictors add little explanatory power for paediatric nutritional status. Future data enrichment (e.g., family diet, adherence, physical activity, household-level SES) may be needed to tease these effects better.

Strengths of study

- Use of spatial-block CV for unbiased evaluation
- Integration of multisource environmental and socioeconomic data
- Temporal external validation to assess transportability
- Reproducible R workflow with version control

Limitations

- Postcode-level exposures may introduce ecological bias
- Residual confounding from unmeasured behavioural factors
- SHAP and permutation importance reflect associations, not necessarily causality

5. Conclusion

This work provides one of the first spatially explicit ML frameworks for predicting CF outcomes in Australia. The model shows strong accuracy, meaningful integration of contextual factors, and practical value for population-health planning.

References

(1) Farinha CM, Callebaut I. Molecular mechanisms of cystic fibrosis – how mutations lead to misfunction and guide therapy. Biosci Rep. 2022 July 1;42(7):BSR20212006.

(2) Fajac I, Burgel PR. Cystic fibrosis. Presse Médicale. 2023 Sept 1;52(3):104169.

(3) Cystic Fibrosis Australia. Australian Cystic Fibrosis Data Registry: 2023 Annual Report [Internet]. Sydney: Cystic Fibrosis Australia; 2024 [cited 2025 June 22]. Available from: https://www.cysticfibrosis.org.au/wp-content/uploads/2024/07/ACFDR_2023_Annual-Report.pdf

(4) Wagner HH. Rethinking the linear regression model for spatial ecological data. Ecology. 2013 Nov;94(11):2381–91.