Exploring Demographic and Environmental Determinants of Rare Pulmonary Disease Using ICEES

A RENCI White Paper
WP-25-08

Brenna Hanson
North Carolina State University, Raleigh, North Carolina, USA

Perry Haaland
Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

Karamarie Fecho
Renaissance Computing Institute, University of North Carolina at Chapel Hill, Chapel Hill, NC, Copperline Professional Solutions, LLC, Pittsboro, North Carolina, USA

How to cite this paper: Hanson, B., Haaland, P., & Fecho, K. (2025). Exploring Demographic and Environmental Determinants of Rare Pulmonary Disease Using ICEES. Renaissance Computing Institute. https://doi.org/10.7921/qyy2-jb07

Corresponding author: Karamarie Fecho, Renaissance Computing Institute, 100 Europa Drive, Suite 540, Chapel Hill, North Carolina, USA; Phone: (919) 445-9640; Email: kfecho@renci.org or kfecho@copperlineprofessionalsolutions.com

The work described herein was supported by the National Center for Advancing Translational Sciences, National Institutes of Health (OT2TR003430)

renci

www.renci.org

Exploring Demographic and Environmental Determinants of Rare Pulmonary Disease Using ICEES

Brenna Hanson[1], Perry Haaland[2], Karamarie Fecho[3,4]

[1]North Carolina State University, Raleigh, NC; [2]Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, NC; [3]Copperline Professional Solutions, LLC, Pittboro, NC; [4]Renaissance Computing Institute, University of North Carolina at Chapel Hill, Chapel Hill, NC

Abstract

The Integrated Clinical and Environmental Exposures Service (ICEES) provides regulatory-compliant open access to semi-aggregated electronic health record data that have been integrated with environmental exposures data. ICEES supports a number of patient data sets, including a data set on patients with rare pulmonary disease (RPD). The RPD data set is composed of patients with suspected or confirmed primary ciliary dyskinesia (PCD), cystic fibrosis (CF), or idiopathic pulmonary bronchiectasis (IB). We utilized the ICEES “multivariate feature analysis” capability within the ICEES RPD endpoint to generate a multivariate feature table for proof-of-concept analysis. We first conducted a bivariate analysis to determine if select demographic factors and environmental exposures differed among patients with a confirmed diagnosis PCD, CF, or IB when compared to patients without a confirmed diagnosis. We found that patients with IB tended to be female; patients with PCD or CF tended to be younger; patients with IB tended to be older; patients with CF or IB tended to be Caucasian and were unlikely to be obese; and patients with CF tended to be exposed to relatively high levels of ozone. We then used the same features to generate a multivariate feature table and apply a multinomial model and a logistic model to further explore the data. We demonstrated an association between a confirmed diagnosis of RPD (PCD, CF, or IB) and relatively high levels of exposure to airborne pollutants (particulate matter and ozone). We discuss our findings, the limitations when working with rare disease data sets, and the value of the ICEES RPD endpoint as a publicly-available resource for researchers.

1. Introduction

Electronic health records (EHRs) provide a rich source of observational real-world data on health and disease. Yet, access to EHR data is restricted to protect patient privacy and governed by federal and institutional regulations such as the US Health Insurance Portability and Accountability Act of 1996[1,2]. Moreover, EHR data typically do not include data on environmental exposures, which are increasingly recognized as important contributors to health and disease[3]. Indeed, exposures to high levels of airborne pollutants, for example, are associated with a wide variety of diseases, including asthma[4], cardiovascular disease[5], cancer[6], diabetes[7], and Alzheimer’s disease[8].

The Integrated Clinical and Environmental Exposures Service (ICEES) was developed to address these challenges. ICEES openly exposes, in a regulatory-compliant manner, EHR data that have been integrated at the patient level with a variety of environmental exposures data[9]. ICEES has been applied to explore the relationship between demographic factors and environmental exposures on asthma and related common pulmonary disorders[10,11,12], primary ciliary dyskinesia (PCD) and related rare pulmonary disorders[13], and several other conditions. ICEES was originally designed to support basic statistical methodologies such as bivariate Chi Square analysis. However, the service was recently extended to support multivariate analysis, albeit with constraints imposed to comply with federal and institutional regulations[14].

Herein, we extend our prior research on PCD and related rare pulmonary disorders[13] by applying bivariate and multivariate statistical models to the ICEES RPD endpoint to explore demographic factors and environmental exposures that are associated with, may predict, and/or may differentiate patients with PCD, cystic fibrosis (CF), or idiopathic pulmonary bronchiectasis (IB).

2. Methods

2.1 Overview of ICEES RPD Data Set

The ICEES RPD data set includes EHR data on patients at UNC Health suspected of having one of three following rare pulmonary diseases, using a complex set of selection criteria[13]: PCD; CF; and IB. A definitive diagnosis of CF was made by way of an affirmative diagnostic code in a given patient’s EHR and/or by manual physician chart review (i.e., consensus by two independent physicians). An EHR diagnostic code does not exist for PCD or IB, so a definitive diagnosis for those rare diseases was made as part of the same physician chart review. Thus, the final RPD data set includes a mix of patients suspected or confirmed to have PCD, CF, or IB (N = 4846).

2.2 Multivariate Feature Table Generation

We selected sex, age, race, obesity, exposure to particulate matter ≤ 2.5-microns in diameter (PM2.5) or ozone for inclusion in our multivariate data set. The choice of features was based primarily on our prior work with the ICEES RPD data set[13].

The ICEES RPD endpoint contains data on multiple study periods (i.e., calendar years 2010-2021). We selected the year 2018 for subsequent analysis, as that year was most enriched in patients with a confirmed diagnosis of PCD, CF, or IB who had one or more annual visits to UNC Health (see Appendix A for details).

Several of the variables in the ICEES RPD data set were unbalanced with some subgroup sizes being too small to provide reliable statistical estimates. We addressed this problem by either dropping categories (and consequently subjects in those categories) that were too small for reliable analysis or by combining (binning) adjacent categories that were too small to analyze independently. (see Appendix A for details). As an example of the first strategy, we restricted the data set based on race to African American (n = 1054) or Caucasian (n = 3792) and dropped from the analysis subjects outside of these categories. Applying the second strategy, we binned Age to 0-17, 18-63, and 64-89 years; PM2.5 exposure to Low (4.94, 8.32 μg/m^3] or High (8.32, 10.57 μg/m^3]; and Ozone to Low (30.26, 43. 96 ppm] or High (43.96, 47.38 ppm].

2.3 Bivariate Analysis Methodology

The purpose of the bivariate analysis was to determine if demographic factors and environmental exposures differed among the three groups of patients (PCD, CF, IB), as differentiated by the variable Diagnosis. To achieve this, we fit a generalized linear model (GLM) between Diagnosis and each feature variable, using Diagnosis as the explanatory variable and the respective feature variable as the response variable. For binary explanatory variables (all except Age), we fit a logistic regression using the glm() function in R. Since Age had three ordered categories, we fit an ordered logistic regression using the polr() function from the MASS package in R. We also analyzed all pairwise comparisons of disease using a Tukey two-sided test from the glht() function of the multcomp package in R.

2.4 Multivariate Analysis Methodology

We further explored the impact of environmental exposures using multivariate analysis methodology. We first created a new variable called Exposure, whose value depended on both PM2.5 exposure and Ozone exposure and their combined level of exposure (i.e., Low PM2.5 and Low Ozone, Low PM2.5 and High Ozone, High PM2.5 and Low Ozone, High PM2.5 and High Ozone, abbreviated as PMLow-OzLow, PMLow-OzHigh, PMHigh-OzLow, and PMHigh-OzHigh, respectively). We conducted a Chi Square test of independence between Exposure and Diagnosis by obtaining a theoretical p-value using the chisq.test() in R and a bootstrapped p-value. We also conducted a Fisher test for independence using a simulated p-value with the fisher.test() function of the stats package in R.

We then fit a multinomial model between Diagnosis and Exposure, with Exposure as the explanatory variable and Diagnosis as the response variable, using the multinom() function in the nnet package in R. This generated multiple pairwise comparisons between the different levels of Exposure and Diagnosis. We assessed the fitted model through z-testing of model coefficients.

Finally, we focused on the highest level of Exposure where both PM2.5 Exposure and Ozone Exposure were relatively high (i.e., PMHigh-OzHigh). To do this, we created an additional variable called High Exposure which took the value of 1 when both PM2.5 Exposure and Ozone Exposure were relatively high and 0 otherwise. We fit a logistic regression between Diagnosis and High Exposure using the glm() function in R, with Diagnosis as the explanatory variable and High Exposure as the response variable. We assessed the fitted logistic regression through z-testing of model coefficients, Analysis of Variance (ANOVA), and predicted probabilities.

3. Results

3.1 Final Multivariate Data Set

The final RPD cohort tended to be female, older, Caucasian, and non-obese with relatively low levels of PM2.5 and ozone exposure. The majority of patients were suspected but not confirmed of having an RPD diagnosis (i.e. no confirmed diagnosis of PCD, CF, or IB), which was expected given the challenge of finding a definitive diagnosis for patients with RPD. Among the patients with a confirmed RPD diagnosis, more patients were diagnosed with CF or IB than PCD, also as expected given that CF and IB are more common than PCD. (Table 1)

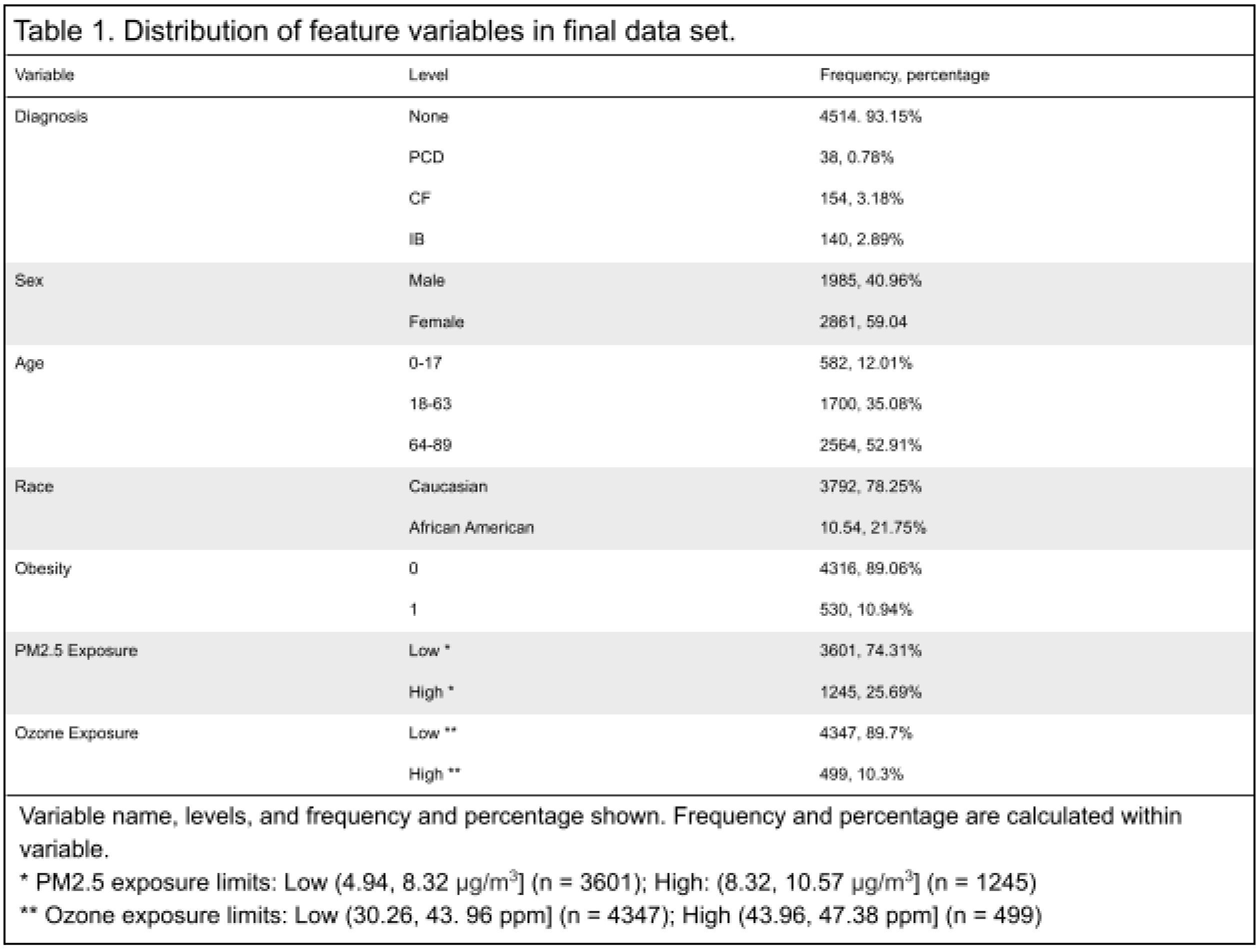

Table 1. Distribution of feature variables in final data set.

Variable	Level	Frequency, percentage
Diagnosis	None	4514. 93.15%
	PCD	38, 0.78%
	CF	154, 3.18%
	IB	140, 2.89%
Sex	Male	1985, 40.96%
	Female	2861, 59.04
Age	0-17	582, 12.01%
	18-63	1700, 35.08%
	64-89	2564, 52.91%
Race	Caucasian	3792, 78.25%
	African American	10.54, 21.75%
Obesity	0	4316, 89.06%
	1	530, 10.94%
PM2.5 Exposure	Low *	3601, 74.31%
	High *	1245, 25.69%
Ozone Exposure	Low **	4347, 89.7%
	High **	499, 10.3%

Variable name, levels, and frequency and percentage shown. Frequency and percentage are calculated within variable.

* PM2.5 exposure limits: Low (4.94, 8.32 μg/m^3] (n = 3601); High: (8.32, 10.57 μg/m^3] (n = 1245)

** Ozone exposure limits: Low (30.26, 43. 96 ppm] (n = 4347); High (43.96, 47.38 ppm] (n = 499)

3.2 Bivariate Analysis Results

We conducted bivariate analyses to determine whether demographic factors and environmental exposures differed among patients with a confirmed PCD, CF, or IB diagnosis when compared to patients suspected to have, but not confirmed to have, RPD.

Patients with a confirmed diagnosis of IB were significantly more likely to be female than patients with confirmed CF or IB, when compared to patients without a confirmed RPD diagnosis (Tukey two-sided p-value <0.001; Figure 1).

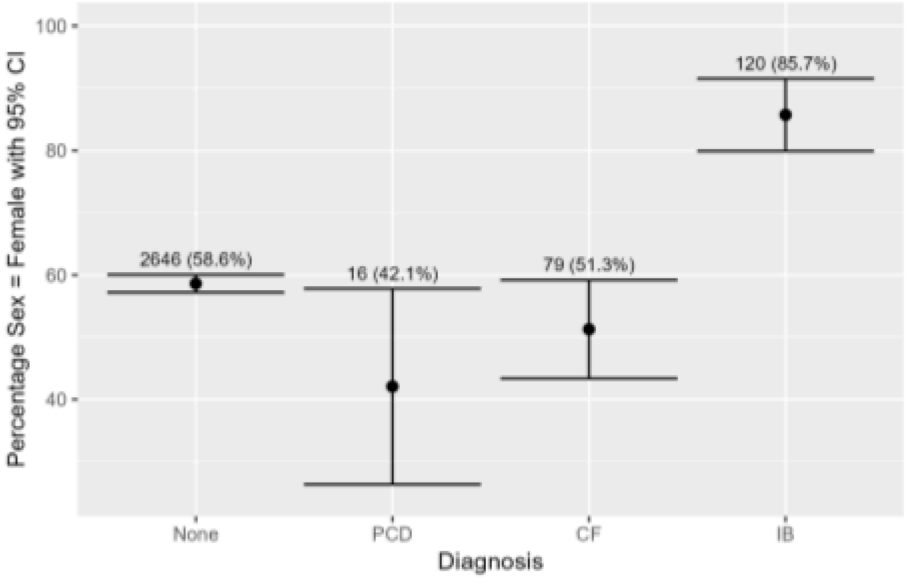

Figure 1. Differences in sex by confirmed diagnosis. Displayed is the predicted percentage of female patients on the y-axis by confirmed diagnosis on the x-axis. A 95% confidence interval with total number and percentage of female patients is shown for each diagnosis group.

Patients with a confirmed diagnosis of PCD (Tukey two-sided p-value < 1e-4) or CF (Tukey two-sided p-value < 1e-4) were more likely to be in the 0-17 years age group, whereas patients with a confirmed IB diagnosis were more likely to be in the 18-63 years age group (Tukey two-sided p-value < 1e-4), when compared to patients without a confirmed RPD diagnosis (Figure 2).

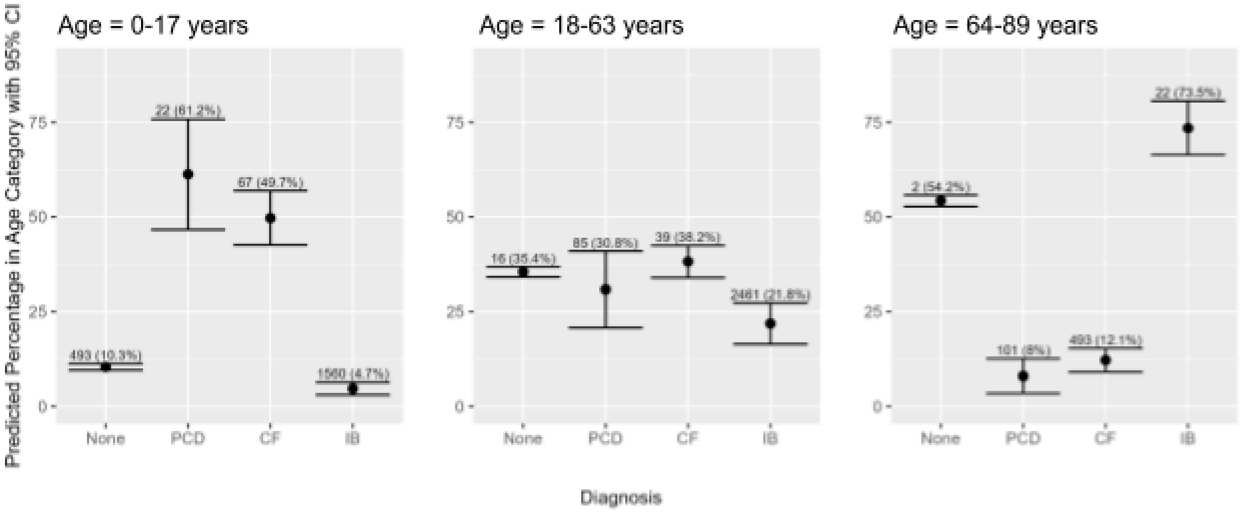

Figure 2. Differences in age by confirmed diagnosis. Displayed is the predicted percentage of patients in age group on the y-axis by confirmed diagnosis on the x-axis. A 95% confidence interval with total number and percentage of patients in age group is shown for each diagnosis group.

Patients with a confirmed RPD diagnosis (CF, PCD, or IB) were more likely to be Caucasian (i.e., less likely to be African American), when compared to patients without a confirmed RPD diagnosis. The difference was found to be significant for patients with a confirmed CF (Tukey two-sided p-value <0.0001) or IB (p-value <0.0001) diagnosis (Figure 3).

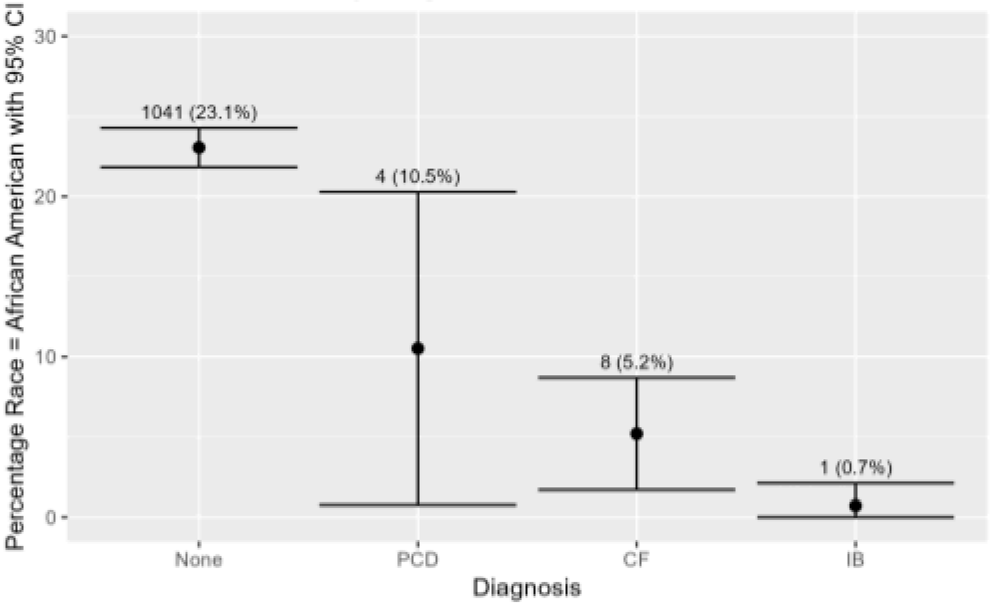

Figure 3. Differences in race by confirmed diagnosis. Displayed is the predicted percentage of African American patients on the y-axis by confirmed diagnosis on the x-axis. A 95% confidence interval with total number and percentage of African American patients is shown for each diagnosis group.

Patients with a confirmed PCD, CF, or IB diagnosis were less likely to be diagnosed with obesity, when compared to those without a confirmed RPD diagnosis. The difference was significant for patients with CF (Tukey two-sided p-value 0.0046) and IB (Tukey two-sided p-value 0.00791) (Figure 4).

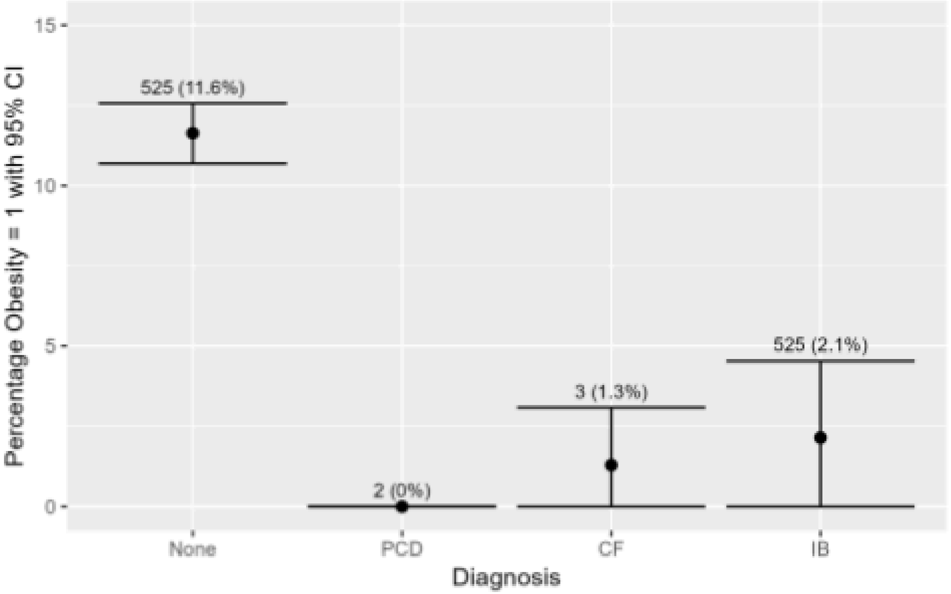

Figure 4. Differences in obesity by confirmed diagnosis. Displayed is the predicted percentage of obese patients on the y-axis by confirmed diagnosis on the x-axis. A 95% confidence interval with total number and percentage of obese patients is shown for each diagnosis group.

The proportion of patients exposed to relatively high levels of PM2.5 was similar among all four patient groups (Figure 5).

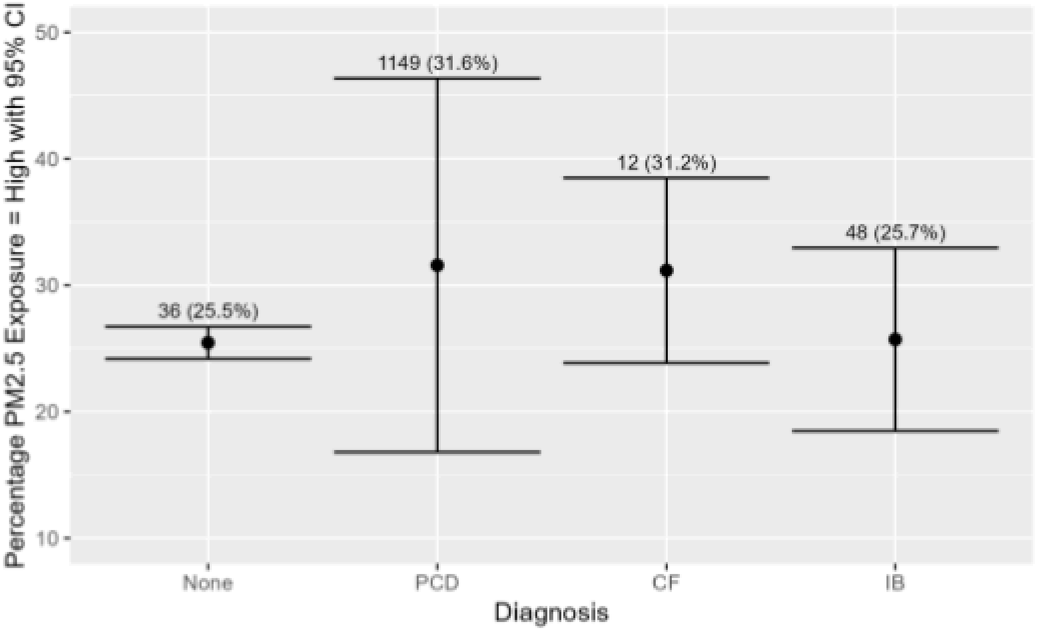

Figure 5. Differences in PM2.5 exposure by confirmed diagnosis. Displayed is the predicted percentage of patients with relatively high PM2.5 exposure on the y-axis by confirmed diagnosis on the x-axis. A 95% confidence interval with total number and percentage of patients with relatively high PM2.5 exposure is shown for each diagnosis group.

The proportion of patients exposed to relatively high levels of ozone was significantly higher among patients with a confirmed CF diagnosis, when compared to those without a confirmed RPD diagnosis (Tukey two-sided p-value <0.001). The proportion of patients exposed to relatively high levels of ozone was similar among patients with a confirmed PCD or IB diagnosis and those without a confirmed RPD diagnosis (Figure 6).

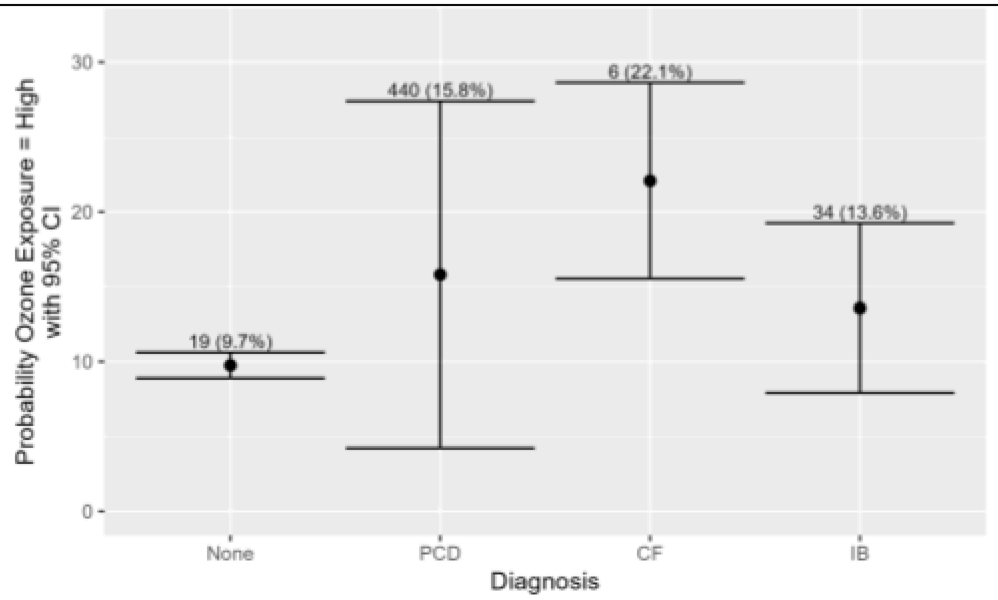

Figure 6. Differences in ozone exposure by confirmed diagnosis. Displayed is the predicted percentage of patients with relatively high ozone exposure on the y-axis by confirmed diagnosis on the x-axis. A 95% confidence interval with total number and percentage of patients with relatively high ozone exposure is shown for each diagnosis group.

3.3 Exposure Risk Analysis

3.3.1 Association of Exposure and Diagnosis

To further explore the relationship between airborne pollutant exposure and confirmed RPD diagnosis, we first examined the overall relationship between Exposure as the explanatory variable and Diagnosis as the response variable. Exposure (PMLow-OzLow, PMLow-OzHigh, PMHigh-OzLow, or PMHigh-OzHigh) and Diagnosis (confirmed PCD diagnosis, confirmed CF diagnosis, confirmed IB diagnosis, or no RPD diagnosis) were found to be significantly associated under the Chi-Square test of independence using both a theoretical (p-value <2.2e-16) and simulated (p-value 0) approach. The Fisher test of independence also yielded a significant result (p-value 0.0005).

3.3.2 Multinomial Model

We then applied a multinomial model to examine the relationship between relative levels of airborne pollutant exposure (Low, High) and Diagnosis. The fitted multinomial model, including coefficient significance, is shown in Table 7. The coefficients comparing the highest level of Exposure (PMHigh-OzHigh) to the lowest level of Exposure (PMLow-OZLow) were significant for all three diagnosed groups, i.e., PCD (p-value 8.137e-06), CF (p-value 8.438e-15), and IB (p-value 1.042e-03). The PCD group was somewhat variable, however, in that the coefficient comparing the PMLow-OzHigh group with the PMLow-OzLow group also was significant (p-value 0).

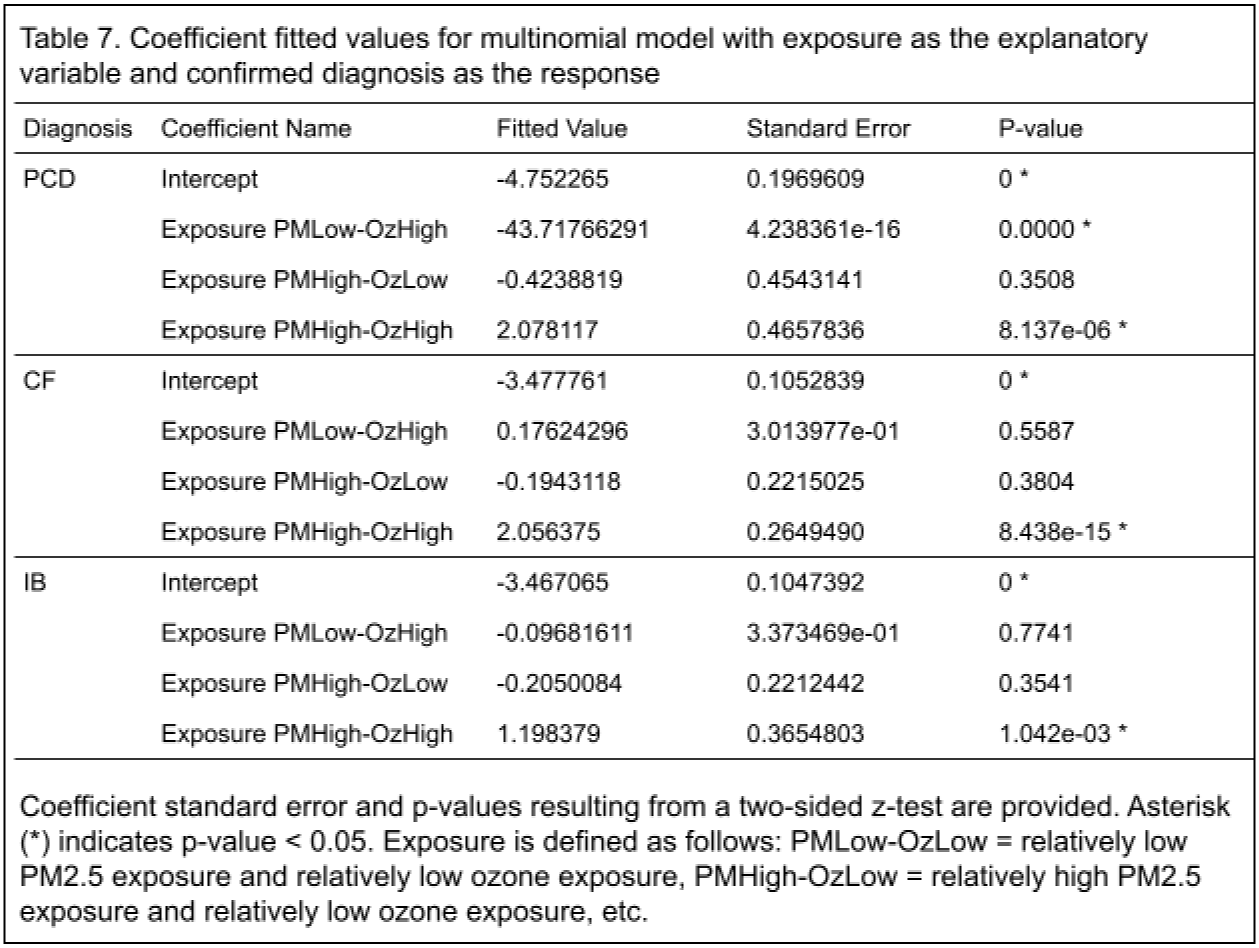

Table 7. Coefficient fitted values for multinomial model with exposure as the explanatory variable and confirmed diagnosis as the response

Diagnosis	Coefficient Name	Fitted Value	Standard Error	P-value
PCD	Intercept	-4.752265	0.1969609	0 *
	Exposure PMLow-OzHigh	-43.71766291	4.238361e-16	0.0000 *
	Exposure PMHigh-OzLow	-0.4238819	0.4543141	0.3508
	Exposure PMHigh-OzHigh	2.078117	0.4657836	8.137e-06 *
CF	Intercept	-3.477761	0.1052839	0 *
	Exposure PMLow-OzHigh	0.17624296	3.013977e-01	0.5587
	Exposure PMHigh-OzLow	-0.1943118	0.2215025	0.3804
	Exposure PMHigh-OzHigh	2.056375	0.2649490	8.438e-15 *
IB	Intercept	-3.467065	0.1047392	0 *
	Exposure PMLow-OzHigh	-0.09681611	3.373469e-01	0.7741
	Exposure PMHigh-OzLow	-0.2050084	0.2212442	0.3541
	Exposure PMHigh-OzHigh	1.198379	0.3654803	1.042e-03 *

Coefficient standard error and p-values resulting from a two-sided z-test are provided. Asterisk (*) indicates p-value < 0.05. Exposure is defined as follows: PMLow-OzLow = relatively low PM2.5 exposure and relatively low ozone exposure, PMHigh-OzLow = relatively high PM2.5 exposure and relatively low ozone exposure, etc.

3.3.3 Logistic Model

Finally, we applied a logistic regression model to examine the relationship between exposure to the highest relative levels of airborne pollutants (i.e. High Exposure) and Diagnosis. As seen in Table 8, the coefficients for the fitted logistic regression between High Exposure (relatively high PM2.5 exposure and relatively high Ozone exposure) and Diagnosis were significant.

Table 8. Coefficient fitted values for logistic regression with confirmed diagnosis for PCD, CF, or IB as the explanatory variable and exposure to high levels of both PM2.5 and ozone as the response variable

Coefficient Name	Fitted Value	Standard Error	P-value
Intercept	-3.9296	0.1083	<2e-16*
Diagnosis PCD	2.2556	0.4579	8.38e-07*
Diagnosis CF	2.0837	0.2586	7.70e-16*
Diagnosis IB	1.2516	0.3612	0.00053*

Coefficient standard error and p-values resulting from a two-sided z-test are provided. Asterisk (*) indicates p-value < 0.05.

Indeed, the proportion of patients exposed to relatively high levels of both PM2.5 and Ozone was higher among patients with a confirmed PCD, CF, and IB diagnosis, when compared to patients without a confirmed RPD diagnosis (Figure 7).

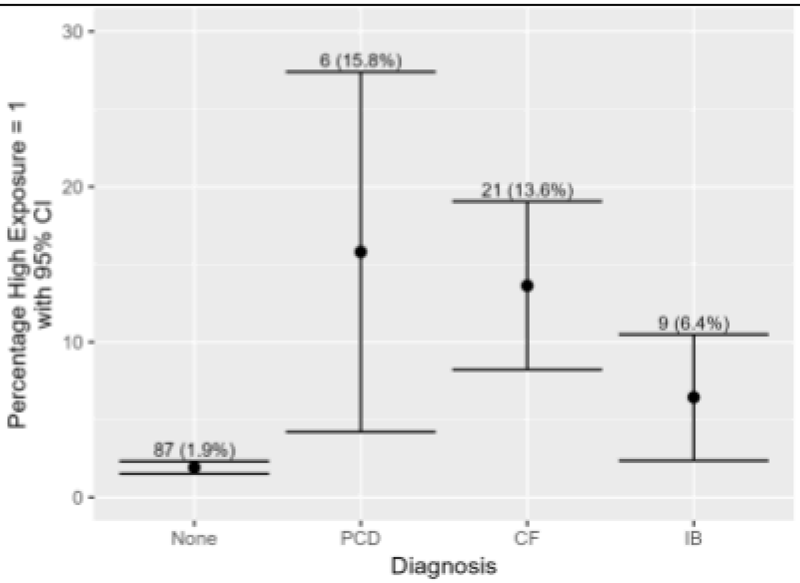

Figure 7. Differences in the percentage of patients with high levels of exposure to both PM2.5 and ozone were analyzed by logistic regression. Displayed on the y-axis is the predicted percentage of patients with high levels of exposure to both PM2.5 and ozone by confirmed diagnosis on the x-axis. A 95% confidence interval with total number and percentage of patients with high levels of exposure to both PM2.5 and ozone is shown for each diagnosis group.

5. Discussion

5.1 Key Findings

Our goal with this study was to leverage the ICEES multivariable feature table functionality at the ICEES RPD endpoint to explore demographic factors and environmental exposures that are associated with, may predict, and/or may differentiate patients with confirmed PCD, CF, and IB. Through bivariate analysis, we investigated differences in the distribution of each feature variable based on confirmed diagnosis. We found that patients with a confirmed IB diagnosis tended to be female, when compared to those without a confirmed RPD diagnosis. Additionally, patients with a confirmed PCD or CF diagnosis tended to be younger, while those with a confirmed IB diagnosis tended to be older, when compared to patients without a confirmed RPD diagnosis. Finally, when compared to patients without a confirmed RPD diagnosis, those with a confirmed CF diagnosis tended to be Caucasian, were not likely to be obese, and had relatively high ozone exposure. Thus, we conclude that sex, age, race, obesity, and ozone exposure exhibit differential distributions based on confirmed RPD diagnosis.

To further explore the impact of environmental exposures, we investigated the relationship between confirmed diagnoses of PCD, CF, or IB and high level of overall airborne pollutant exposure (i.e., PM2.5 and ozone). We found a strong relationship between the three confirmed diagnoses and overall airborne pollutant exposure, but only at the highest level of exposure (i.e., relatively high levels of both PM2.5 and ozone). Thus, we conclude that patients with a confirmed diagnosis for PCD, CF, or IB tend to be exposed to relatively high levels of PM2.5 and ozone in our RPD data set. The relationship between airborne pollutant exposure and confirmed RPD diagnosis is unlikely to be causal; rather, we speculate that high levels of airborne pollutant exposure may increase the severity of disease thereby making a confirmed diagnosis more likely.

5.2 Limitations

First, we note that data loss is inherent to the ICEES multivariate table functionality in order to preserve regulatory compliance. This creates the possibility of bias in the retrieved data set and may degrade model quality[14]. However, in this study, the proportion of patients with confirmed PCD, CF, or IB was similar before and after data multivariate table generation, which provides confidence in the resultant data set (see Appendix A).

Second, our analysis was limited by the imbalance inherent in the RPD data set, which impeded our ability to apply a robust predictive analysis. This is a common problem when investigating rare disease. To account for the imbalance within our data set, we carefully examined the structure of the data and dropped or combined sparse categories where the imbalance was too great to provide reliable statistical estimates. This introduced data loss and reduced data granularity. We are exploring additional methods to account for imbalance.

5.2 Conclusion

Our primary goal with the work reported herein was to develop a proof-of-concept analysis of rare pulmonary disease utilizing the ICEES multivariate table creation capability to explore demographic factors and environmental exposures among patients with a confirmed diagnosis of PCD, CF, or IB for comparison with patients with a suspected but not confirmed diagnosis of rare pulmonary disease. Several demographic factors and environmental exposures emerged as worthy of consideration in future studies using the ICEES RPD data set. As such, we note the value of the ICEES RPD data set as a unique open source of integrated clinical and environmental exposures data on rare pulmonary disease for in-depth analysis and exploration of demographic factors and environmental determinants that might influence disease severity and health outcomes.

Funding Support

The work described herein was supported by the National Center for Advancing Translational Sciences, National Institutes of Health (OT2TR003430).

Acknowledgements

The authors gratefully acknowledge the expertise provided by Dr. David B. Peden, Dr. Michael R. Knowles, and Dr. Margaret Leigh at the UNC School of Medicine. The authors also appreciate the support provided by Dr. Ashok Krishnamurthy, Director of the Renaissance Computing Institute.

References

1. (OCR) O for CR. Summary of the HIPAA privacy rule [Internet]. Department of Health and Human Services; 2025 [cited 2025 Mar 31]. Available from: https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html

2. Ahalt SC, Chute CG, Fecho K, et al. Clinical Data: Sources and Types, Regulatory Constraints, Applications. Clin Transl Sci. 2019;12(4):329-333. doi:10.1111/cts.12638

3. Wu H, Eckhardt CM, Baccarelli AA. Molecular mechanisms of environmental exposures and human disease. Nat Rev Genet. 2023;24(5):332-344. doi:10.1038/s41576-022-00569-3

4. Aghapour M, Ubags ND, Bruder D, et al. Role of air pollutants in airway epithelial barrier dysfunction in asthma and COPD. Eur Respir Rev. 2022;31(163):210112. Published 2022 Mar 23. doi:10.1183/16000617.0112-2021

5. Miller MR. The cardiovascular effects of air pollution: Prevention and reversal by pharmacological agents. Pharmacol Ther. 2022;232:107996. doi:10.1016/j.pharmthera.2021.107996

6. Lagunas-Rangel FA, Liu W, Schiöth HB. Can Exposure to Environmental Pollutants Be Associated with Less Effective Chemotherapy in Cancer Patients?. Int J Environ Res Public Health. 2022;19(4):2064. Published 2022 Feb 12. doi:10.3390/ijerph190420644

7. Khalil WJ, Akeblersane M, Khan AS, Moin ASM, Butler AE. Environmental Pollution and the Risk of Developing Metabolic Disorders: Obesity and Diabetes. Int J Mol Sci. 2023;24(10):8870. Published 2023 May 17. doi:10.3390/ijms24108870

8. Olloquequi J, Díaz-Peña R, Verdaguer E, Ettcheto M, Auladell C, Camins A. From Inhalation to Neurodegeneration: Air Pollution as a Modifiable Risk Factor for Alzheimer's Disease. Int J Mol Sci. 2024;25(13):6928. Published 2024 Jun 25. doi:10.3390/ijms25136928

9. Fecho K, Pfaff E, Xu H, Champion J, Cox S, Stillwell L, Bizon C, Peden D, Krishnamurthy A, Tropsha A, Ahalt SC. A novel approach for exposing and sharing clinical data: the Translator Integrated Clinical and Environmental Exposures Service. J Am Med Inform Assoc 2019;26(10):1064–1073. doi: 10.1093/jamia/ocz042.

10. Fecho K,* Ahalt SC, Appold S, Arunachalam S, Pfaff E, Stillwell L, Valencia A, Xu H, Peden D. Development and application of an open tool for sharing and analyzing integrated clinical and environmental exposures data: asthma use case. JMIR Form Res 2022;6(4):e32357. doi: 10.2196/32357. *Apart from first/lead and last/senior author, all other authors are listed in alphabetical order.

11. Xu H, Cox S, Stillwell L, Pfaff E, Champion J, Ahalt SC, Fecho K. FHIR PIT: an open software application for spatiotemporal integration of clinical data and environmental exposures data. BMC Med Inform Decis Mak 2020;20:article 53. doi: 10.21203/rs.2.19633/v1.

12. Fecho K,* Ahalt S, Arunachalum S, Champion J, Chute CG, Gersing K, Glusman G, Hadlock J, Lee J, Pfaff E, Robinson M, Sid E, Ta C, Xu H, Zhu R, Zhu Q, Peden DB, and The Biomedical Data Translator Consortium. Sex, obesity, diabetes, and exposure to particulate matter: scientific insights revealed by analysis of open clinical data sources during a five-day hackathon. J Biomed Inform 2019;100:103325 [Special Communication]. doi: 10.1016/j.jbi.2019.103325. *Apart from first/lead and last/senior author, all other authors are listed in alphabetical order.

13. Fecho K*, Ahalt SC, Knowles M, Krishnamurthy A, Leigh M, Morton K, Pfaff E, Wang M, Yi H. Leveraging open electronic health record data and environmental exposures data to derive insights into rare pulmonary disease. Front Artif Intell 2022; 5:918888 (special issue on Biomedical Informatics Applications in Rare Diseases). doi: 10.3389/frai.2022.918888. *Apart from the first author, all authors are listed in alphabetical order.

14. Fecho K,* Haaland P, Krishnamurthy A, Lan B, Ramsey S, Schmitt PL, Sharma P, Sinha M, Xu H. An approach for open multivariate analysis of integrated clinical and environmental exposures data. Inform Med Unlocked 2021;26:100733. doi.org/10.1016/j.imu.2021.100733. *Apart from first/lead author, all other authors are listed in alphabetical order.

15. Schmitt, P. L., Fecho, K., Haaland, P., Krishnamurthy, A., Lan, B., Sharma, P., & Sinha, M. A Framework for Estimating the Bounds of Contingency Tables: Application to an Open Clinical Research Service. RENCI Technical Report, TR-22-01, Renaissance Computing Institute, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. doi: 10.7921/9ea1-h198. https://renci.org/technical-reports/tr-22-01.

Appendix A

Here, we describe details on the ICEES RPD data set and the methodology we used to generate and evaluate the multivariate table that was used for subsequent analysis.

A.1 Overview of ICEES RPD Data Set

Our general goal was to develop and apply a multivariate analytic model to the ICEES RPD data set. The ICEES RPD data set includes EHR data on patients at UNC Health suspected of having one of three following rare pulmonary diseases, using a complex set of selection criteria: PCD; CF; or IB. A definitive diagnosis of CF was made by way of an affirmative diagnostic code in a given patient's EHR and/or by manual physician chart review (i.e., consensus by two independent physicians). An EHR diagnostic code does not exist for PCD or IB, so a definitive diagnosis for those rare diseases was made as part of the same physician chart review. Thus, the final RPD data set includes a mix of patients suspected or confirmed to have PCD, CF, or IB.

A.2 Feature Selection

We selected sex, age, race, obesity, exposure to particulate matter ≤ 2.5-microns in diameter (PM2.5) or ozone for inclusion in our multivariate data set. The choice of features was based primarily on our prior work with the ICEES RPD data set[13]. As shown in Table 1, all feature variables were categorical.

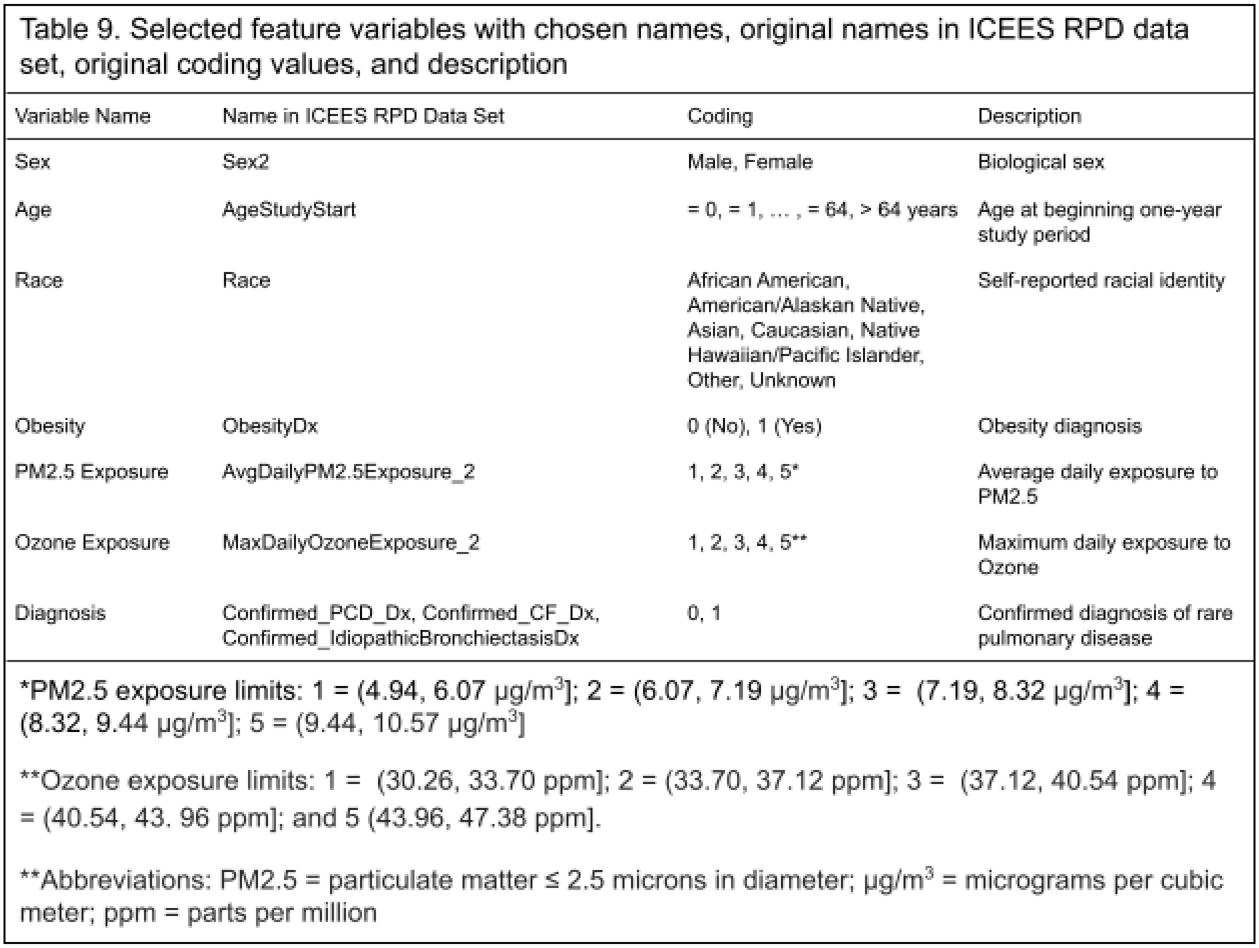

Table 9. Selected feature variables with chosen names, original names in ICEES RPD data set, original coding values, and description

Variable Name	Name in ICEES RPD Data Set	Coding	Description
Sex	Sex2	Male, Female	Biological sex
Age	AgeStudyStart	= 0, = 1, … , = 64, > 64 years	Age at beginning one-year study period
Race	Race	African American, American/Alaskan Native, Asian, Caucasian, Native Hawaiian/Pacific Islander, Other, Unknown	Self-reported racial identity
Obesity	ObesityDx	0 (No), 1 (Yes)	Obesity diagnosis
PM2.5 Exposure	AvgDailyPM2.5Exposure_2	1, 2, 3, 4, 5*	Average daily exposure to PM2.5
Ozone Exposure	MaxDailyOzoneExposure_2	1, 2, 3, 4, 5**	Maximum daily exposure to Ozone
Diagnosis	Confirmed_PCD_Dx, Confirmed_CF_Dx, Confirmed_IdiopathicBronchiectasisDx	0, 1	Confirmed diagnosis of rare pulmonary disease

*PM2.5 exposure limits: 1 = (4.94, 6.07 μg/m^3]; 2 = (6.07, 7.19 μg/m^3]; 3 = (7.19, 8.32 μg/m^3]; 4 = (8.32, 9.44 μg/m^3]; 5 = (9.44, 10.57 μg/m^3]

**Ozone exposure limits: 1 = (30.26, 33.70 ppm]; 2 = (33.70, 37.12 ppm]; 3 = (37.12, 40.54 ppm]; 4 = (40.54, 43. 96 ppm]; and 5 (43.96, 47.38 ppm].

**Abbreviations: PM2.5 = particulate matter ≤ 2.5 microns in diameter; μg/m^3 = micrograms per cubic meter; ppm = parts per million

A.3 Year Selection

The ICEES RPD data set contains data across multiple study periods (i.e., calendar years). The frequency of patients with a confirmed diagnosis of PCD, CF, or IB will vary across years based on whether or not a given patient visits UNC Health in any given year. Thus, we analyzed the frequency of patients with a confirmed diagnosis of PCD, CF, or IB by year, with the goal of selecting a year for subsequent study that was most enriched for patients with a confirmed diagnosis of PCD, CF, or IB (Figure 1). We selected the year 2018 for subsequent analysis, as that year was most enriched in patients for all three RPD diagnoses: PCD, CF, and IB.

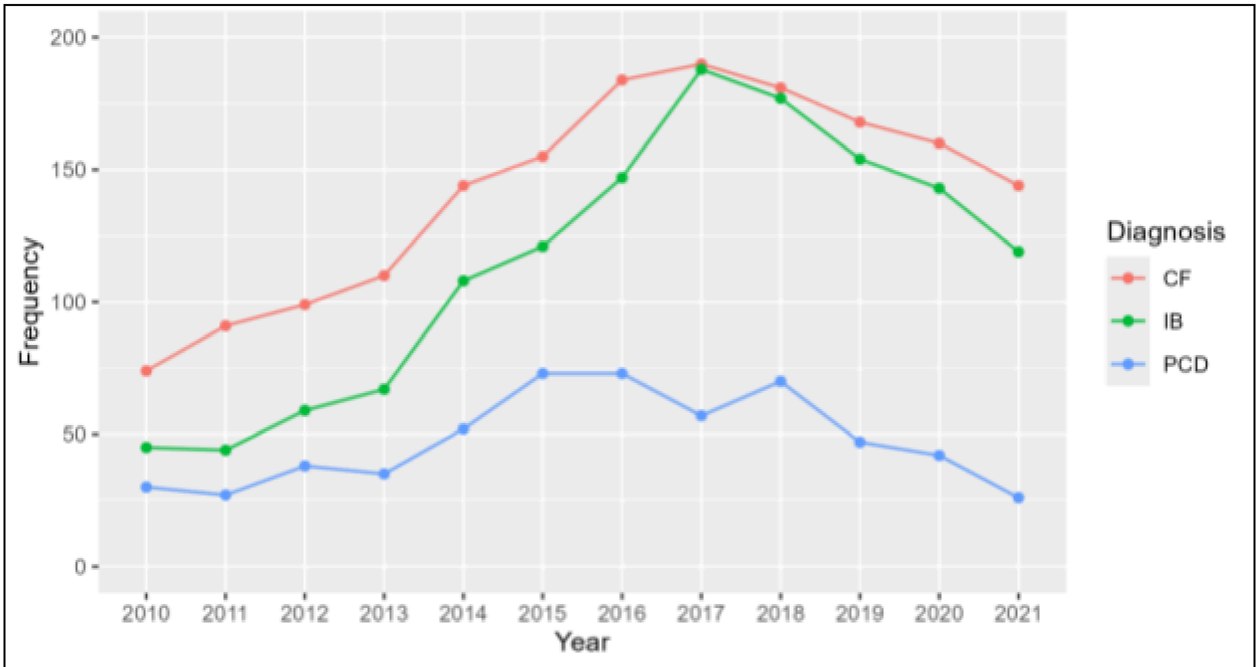

Figure 8. Frequency of patients in the ICEES RPD data set with a positive confirmed diagnosis by year. Grouped by diagnosis, as defined in Table 1.

A.4 Multivariate Table Generation

Based on our selected feature variables and study period year, we then extracted four mutually exclusive data sets using the ‘Multivariate Feature Analysis’ function available at the ICEES RPD OpenAPI endpoint: one for patients with a confirmed diagnosis of PCD; one for patients with a confirmed diagnosis of CF; one with patients with a confirmed diagnosis of IB; and one for the remaining patients who were suspected but not confirmed to have a rare pulmonary disease. These tables were read in using a custom function we created named readMultivariateTable(). We then collapsed the four tables into one and formatted them to be patient-level using dplyr.

A.5 Assessment of Data Loss

By nature of the ICEES multivariate algorithm, which was designed to support the generation of multivariate feature tables while preserving patient anonymity and abiding by all regulatory restrictions, some level of data loss may be introduced into the produced multivariate feature tables[15]. We assessed data loss in the four extracted data sets and in the combined data set by comparing the total cohort size in 2018 to the cohort sizes for the four selected data sets for each diagnosis and for the combined data set.

We found that data loss was generally proportional to cohort size but acceptably minimal for each data set (ranging from 27-1204 patients), thus justifying further analysis, albeit with imbalances in each data set (Table 10).

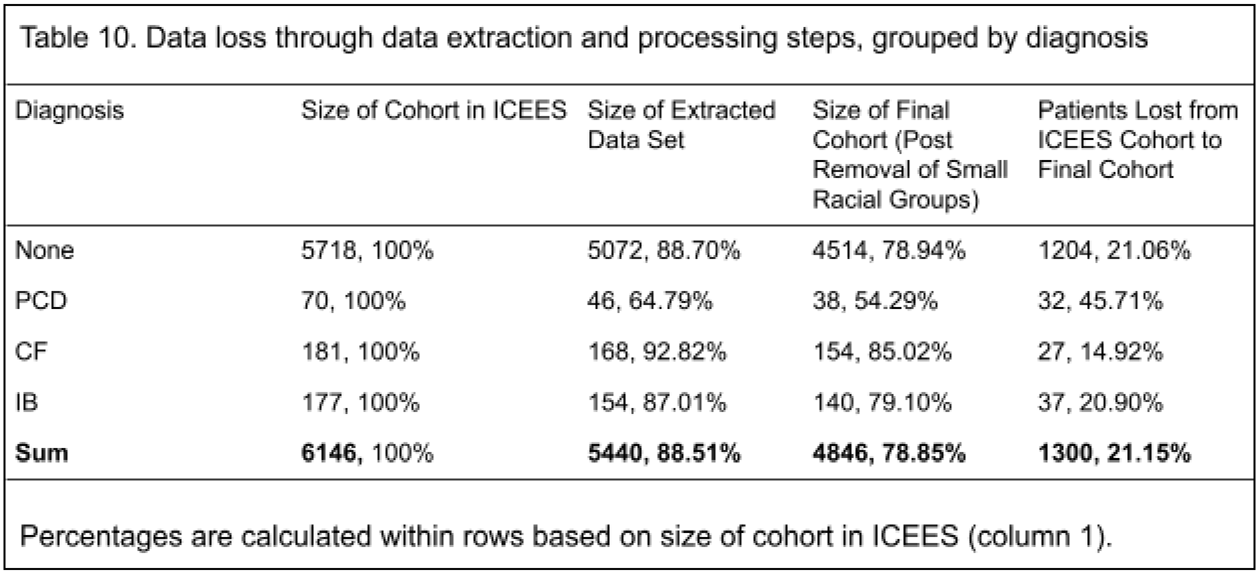

Table 10. Data loss through data extraction and processing steps, grouped by diagnosis

Diagnosis	Size of Cohort in ICEES	Size of Extracted Data Set	Size of Final Cohort (Post Removal of Small Racial Groups)	Patients Lost from ICEES Cohort to Final Cohort
None	5718, 100%	5072, 88.70%	4514, 78.94%	1204, 21.06%
PCD	70, 100%	46, 64.79%	38, 54.29%	32, 45.71%
CF	181, 100%	168, 92.82%	154, 85.02%	27, 14.92%
IB	177, 100%	154, 87.01%	140, 79.10%	37, 20.90%
Sum	**6146,** 100%	**5440, 88.51%**	**4846, 78.85%**	**1300, 21.15%**

Percentages are calculated within rows based on size of cohort in ICEES (column 1).

A.6 Variable Binning

Several of the variables in the ICEES RPD data set were unbalanced and required adjustments to support reliable statistical analysis. We did this by either dropping distinct categories with sample sizes that were too small or by combining (rebinning) adjacent categories to achieve better balance. For example, the race variable was originally distributed as follows: African American (n=1054); American/Alaskan Native (n=52); Asian (n=86); Caucasian (n=3792); Native Hawaiian/Pacific Islander (n=7); Other (2131-1) (n=339); and Unknown (n=110). Because race is a potentially important explanatory variable, it did not seem reasonable to combine small race categories, so we decided to retain only those racial categories that were large enough for reliable analysis. Thus, we restricted race to African American or Caucasian, which decreased the total cohort size from 5,440 patients to 4,846 patients. We also adjusted three other variables by combining adjacent categories into groups that were coherent and also large enough to provide reliable statistical estimates. Age was originally coded as =0, =1, =2, … , =64, >64 years. We collapsed Age into three bins of reasonable size: 1. [0, 17] (n = 582); 2. [18, 63] (n = 1700); and 3. [64, 89] (n= 2564). PM2.5 exposure was distributed as: 1. (4.94, 6.07 μg/m^3] (n = 222); 2. (6.07, 7.19 μg/m^3] (n = 685); 3. (7.19, 8.32 μg/m^3] (n = 2694); 4. (8.32, 9.44 μg/m^3] (n = 1237); 5. (9.44, 10.57 μg/m^3] (n = 8). We collapsed PM2.5 exposure into two bins: Low (4.94, 8.32 μg/m^3] (n = 3601); High: (8.32, 10.57 μg/m^3] (n = 1245). Ozone exposure was distributed as: 1. (30.26, 33.70 ppm] (n = 1); 2. (33.70, 37.12 ppm] (n = 1); 3. (37.12, 40.54 ppm] (n = 33); 4. (40.54, 43. 96 ppm] (n = 4312); and 5. (43.96, 47.38 ppm] (n = 499). We collapsed ozone exposure into two bins: Low (30.26, 43. 96 ppm] (n = 4347); High (43.96, 47.38 ppm] (n = 499).

A.7 Final ICEES RPD Data Set

The distribution of all feature variables in the finalized data set is shown in Table 1 of the main manuscript. Briefly, the final cohort tended to be female, older, Caucasian, and non-obese with low levels of PM2.5 and ozone exposure. The majority of patients were classified as Diagnosis = None, which was expected given the focus on rare disease and the challenge of finding a definitive diagnosis for patients with rare pulmonary disease. Among the patients with a confirmed diagnosis, more patients were diagnosed with CF or IB than PCD, also as expected, given that CF and IB are more common than PCD.