The Impact of Demographic and Clinical Factors on Lung Disease Development and Recovery Outcomes

Mingyu Li, Jack Lohse, Marija Mitevska, Devin DuBois

Abstract

This study investigates how demographic and clinical factors impact lung disease diagnosis, hospital utilization, and recovery outcomes. Using an observational dataset of 5,200 patients, we applied ANOVA, Chi-square tests, and logistic regression models to evaluate associations between smoking status, age, gender, lung capacity, and recovery. Smoking status was not significantly associated with hospital visits or recovery. Multinomial logistic regression found no strong link between age or gender and disease type. Lung capacity also did not significantly influence hospital visitation frequency. Finally, logistic regression models, including stepwise interaction terms, showed no significant predictors of recovery. Overall, none of the tested variables reached statistical significance. These findings suggest that other factors not captured in the dataset, such as disease severity, treatment adherence, or socioeconomic status, may better explain recovery outcomes. Future studies using longitudinal or clinically detailed data are needed to explore these effects further.

Introduction

Chronic respiratory illnesses, most notably COPD, bronchitis, pneumonia, and cancer of the lungs, are leading causes of worldwide disease burden. COPD alone causes over three million annual deaths and is regarded as the third leading cause of death globally. Increases in incidence of these conditions, most notably among elderly populations and in low- and middle-income nations, have led public health professionals and clinicians to examine their etiology and course. While certain risks are well defined, i.e., smoking and environmental exposure, roles for demography, i.e., age and sex, and physiological measures, i.e., lung capacity, have been less well defined in actual world settings.

Lung capacity, or the air volume an individual can expel upon deep breathing, is a direct physiological measure of pulmonary function and one employed frequently to quantify disease severity in clinical environments. Its prospective predictive power for downstream events like hospital use or subsequent recovery, however, is not as well defined at the population level. Similarly, whereas a generally accepted risk for developing chronic disease with increasing age exists, the specific influence of age upon various disease endpoints of the lungs is multifaceted. Gender differences have been noted through previous studies, including underdiagnosis of COPD specifically in females, resulting from historically male-dominant diagnostic models. There is an established association between smoking and lung health, and this study tests to see if smoking status, as documented in this dataset, is linked with actual healthcare use patterns and recovery status. Previous research has generally targeted clinical samples or prospective studies with more controlled measurements. The study here supplements that research by examining those associations within a cross-sectional observational dataset, with an intention of ascertaining which factors, if at all, will be predictors of diagnosis, utilization, and recovery if taken into account individually.

This study explores the question: What lifestyle or biological factors influence disease severity, healthcare utilization, or recovery outcomes among patients with lung disease? To address this, four specific research questions guide the analysis: whether COPD diagnosis can be predicted by age and sex; whether hospitalization rates are influenced by lung capacity; whether hospitalization, with or without adjustment for lung capacity, predicts recovery status; and whether hospital utilization and recovery differ by smoking status. These questions are investigated using well-established statistical methods applied to a dataset large enough to support subgroup analyses and interaction effects. This approach allows for a comprehensive evaluation of how known risk factors contribute to variation in lung health outcomes within an observational framework.

Methods

Data Collection

Data for this study was drawn from an open-source dataset available at Kaggle, entitled “Lungs Diseases Dataset” (Dalvi, 2023). The dataset contains 5,200 observations of persons with different statuses of lung health, including demographics, medical diagnosis, and treatment results. No indication of how the data were obtained is provided through the original documentation, whether through hospital records, surveys, or clinical trials. It seems, however, to have been collated for research or educational purposes. Records contain fields including age, gender, smoking status, lung capacity, hospital admissions, disease type, type of treatment, and state of recovery. As it is a secondary dataset, non-response details and eligibility rates are lacking, and representativeness to the larger populace cannot be ascertained. We assume the dataset to be cross-sectional and observational.

Contributions

We divided the general research question into four subsections, with each team member responsible for one part of the analysis. All of us contributed to writing the initial proposal. Mingyu and Marija were responsible for selecting the dataset, while Mingyu and Jack worked together to come up with the overall research question. The remaining sections of the project were carried out by Mingyu.

Variable Creation

We used several variables to answer each group member’s question. To look at how age and gender affect the chance of having a disease, we used Age as a number that goes up or down and Gender as a group variable with two choices: Male or Female. To see how smoking status affects hospital visits and recovery, we changed Smoking Status into a yes-or-no group, and we made a new variable called Recovered_binary. This was set to 1 for “Yes” and 0 for “No.”To look at how lung capacity and hospital visits are related, we used Lung Capacity as a number and Hospital Visits as a count of healthcare use. For questions about how hospital visits affect getting better or not, we used both Hospital Visits and Recovered_binary. We also made a new yes-or-no variable called has_COPD. It was set to 1 if the Disease Type was “COPD” and 0 if it was not. This helped us focus on people with that lung disease.

Analytic Methods

To determine whether smoking status has an impact on health outcomes, two statistical tests were utilized. A one-way ANOVA was used to test for differences in the average number of hospital visits in three categories of smoking status: smokers, non-smokers, and missing smoking data. This enabled us to examine whether smoking is related to increased healthcare utilization. Second, in order to determine whether recovery outcome differed by smoking status, a Chi-square test of independence was used. This tested whether recovery rates significantly differed by smoking status.

To determine whether age and gender have a relationship to the type of lung disease an individual is diagnosed with, we used a multinomial logistic regression model. This model is appropriate when the outcome variable is categorical with more than two categories. The dependent variable in this case was Disease Type, which was composed of categories such as COPD, Asthma, Bronchitis, Lung Cancer, and Pneumonia. The independent variables were Age and Gender. Using nnet::multinom() function, we estimated the probability of each disease type occurring based on age and gender, and visualized the predicted probabilities across age ranges for both males and females.

After the development of lung disease, the lung capacity of individuals varies. This variation of lung capacity leads us to the question of whether the frequency of hospital visits is affected by the diminishment of lung capacity. To determine this, a One-Way Analysis of Variance (ANOVA) is conducted to determine whether there is a statistically significant difference in the frequency of hospital visits among individuals with differing lung capacity.

To answer the question of whether hospital visits, along with lung capacity, disease type, or treatment type, affect recovery, we used a logistic regression model. We added variables one at a time using stepwise selection, with each variable tested in combination with hospital visits. This helped us build a model that included the most useful predictors while keeping the model simple.

Results

Smoking and Hospital Visits

The dataset contained 5,200 observations of smoking status, recovery status, and hospital visit numbers. 4,617 cases were available for use in hospital visit analysis after removing missing values. A boxplot of hospital visit numbers by smoking status is displayed in Fig. 1.

From the boxplot, distributions of hospital visits were generally consistent across smoking categories. Both smokers and non-smokers were found to have similar medians and interquartiles. Each group's variability was high, and no significant visual differences were apparent. and N/A group is added for completeness of the dataset, and will not be used in the analysis.

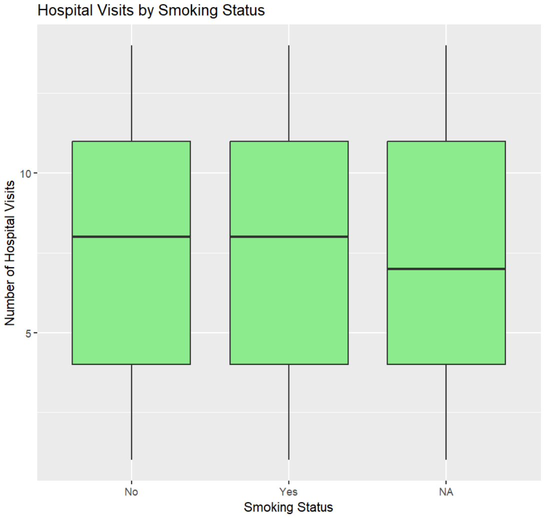

Figure 1. *Number of Hospital Visits by Smoking Status*.
This boxplot displays the distribution of hospital visits across three smoking status categories: Non-smokers ("No") and smokers ("Yes"). The median number of visits and the interquartile ranges appear similar across groups, indicating no substantial difference in hospital usage based on smoking behavior.

One-way Analysis of Variance was used to see whether, based on smoking status, differences in hospital visitation were statistically significant. It was found that there were no significant differences within smoking categories (excluding the NA group) in terms of hospital visits. F(1,4616)=0.084, p=0.772. This indicates that smoking status was not found to be an independent predictor of healthcare use, as quantified in terms of hospital utilization in this data.

Figure 2 is a stacked bar graph of recovery status (recovered, not recovered, NA) by smoking status. In each of the two categories (recovered and not recovered), the proportion of individuals who recovered was just over 50%, without a visible or considerable difference. The NA group will not be considered in the analysis. The proportion who were "not recovered" was almost identical in smokers and non-smokers, and the missing recovery status was low.

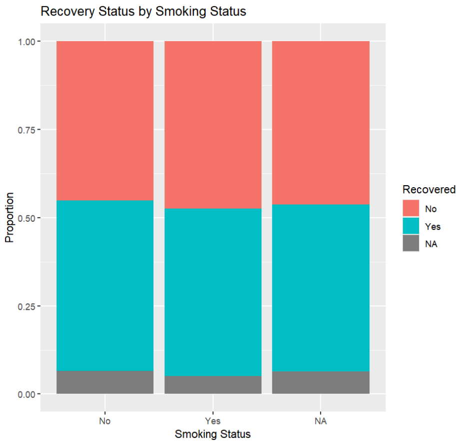

Figure 2. *Recovery Status Proportions by Smoking Status*
This stacked bar chart shows the proportion of patients who recovered, did not recover, or had missing recovery data within each smoking status group. The distribution of recovery outcomes appears similar across smokers and non-smokers, suggesting no clear association between smoking behavior and recovery status.

To determine whether smoking status was related to recovery status, a Chi-square test of independence was used. The outcome of this test was not found to be statistically significant. χ2 = 1.2312, df = 1, p = 0.2672, suggesting that recovery status was unrelated to smoking status in this sample.

Age, Gender, and Disease Type

This study also sought to examine how demographic factors — specifically age and gender — relate to the type of lung disease diagnosed in individuals. The dataset included 5,200 observations, all of which had a lung disease diagnosis. After removing records with missing age, gender, or disease type information, 4,895 cases were included in the analysis.
Figure 3 displays density plots of patient age by disease type, separated by gender. While most disease types were distributed across a broad age range, a few patterns were visible. Asthma and bronchitis appeared more frequently in younger individuals, whereas pneumonia skewed toward

older patients. Gender differences were subtle; however, females with pneumonia trended younger than their male counterparts.

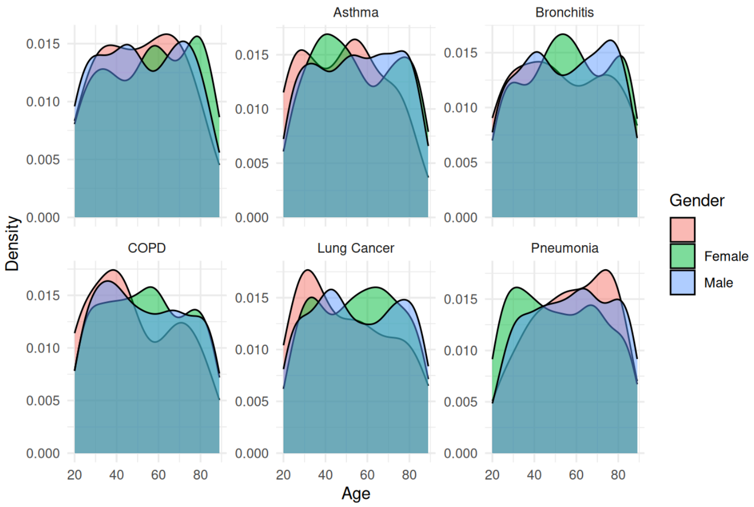

Figure 3. Age Distribution by Disease Type and Gender
This figure displays kernel density plots for age distributions within each lung disease category, separated by gender. Although distributions overlap significantly, certain conditions like asthma appear more common in younger patients, and Pneumonia is more common in older individuals.

To assess whether age and gender influence the type of lung disease diagnosed, a multinomial logistic regression was conducted. The model used disease type as the outcome variable and included age and gender as predictors. The results indicated that neither age or gender was a statistically significant predictor of disease type. For example, the p-value for age in predicting COPD was 0.725, and for gender, the p-values were above 0.1 in nearly all disease categories. One exception was pneumonia, where males had a statistically lower probability of diagnosis compared to females ($p = 0.018$).

A second model was run with an interaction term between age and gender to assess whether the relationship between age and disease diagnosis differed by gender. This interaction was also not statistically significant ($p > 0.4$ across all interaction terms), suggesting that the effect of age on disease diagnosis does not vary by gender.

Lastly, age was categorized into three groups (20–40, 41–60, 61+) to check for nonlinear effects. This alternative model produced results consistent with the continuous age model; again, no significant differences were found between age groups for any specific disease type.

Figure 4 shows the predicted probability of each disease by age and gender based on the regression model. The lines were nearly flat, indicating no strong age-related trends in predicted probability. Similarly, male and female lines closely overlapped, supporting the regression results that gender has little influence on disease type in this dataset.

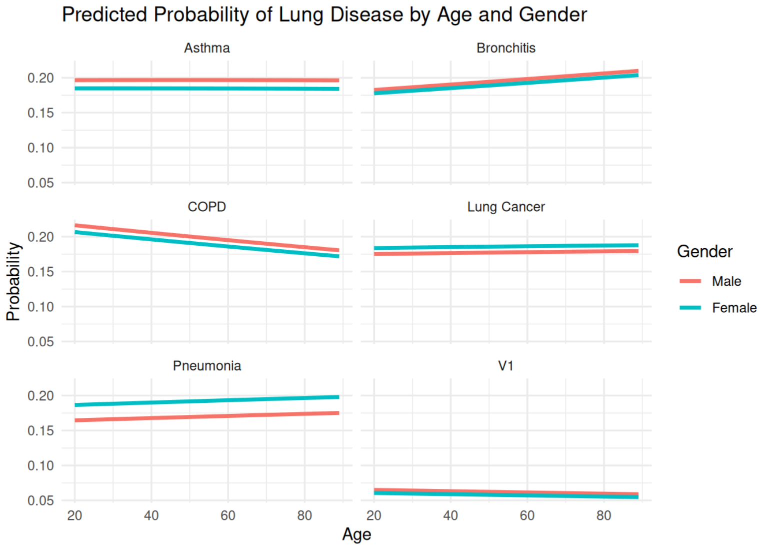

Figure 4. Predicted Probability of Lung Disease by Age and Gender
This line chart shows model-estimated probabilities for each lung disease category across the age spectrum, separated by gender. Lines remain mostly parallel and flat, indicating that age and gender do not meaningfully affect diagnosis probabilities.

Together, these results indicate that while age and gender show some visual differences in lung disease distribution, these patterns are not statistically significant. The likelihood of being diagnosed with a specific type of lung disease does not vary meaningfully by age or gender within this dataset. These results remained consistent across several approaches — using continuous age, age categories, and interaction terms between age and gender — indicating the strength of the primary conclusion.

Lung Capacity and Visits

Lung capacity is the volume of air that an individual's lungs are able to inhale. For a healthy adult, the average lung capacity is approximately 6.00 liters. The data of 5,200 individuals with lung disease was collected, from which 4,615 had both the number of hospital visits and their lung capacity. The lung capacity of the individuals ranges from 1.00 to 6.00 liters.

The histogram (Figure 5) displays the mean number of hospital visits related to lung capacity. Among the range of lung capacity, the mean number of hospital visits falls between 6 and 8 visits. While there is some variation, there is no clear trend demonstrating a relationship between lung capacity and hospital visits.

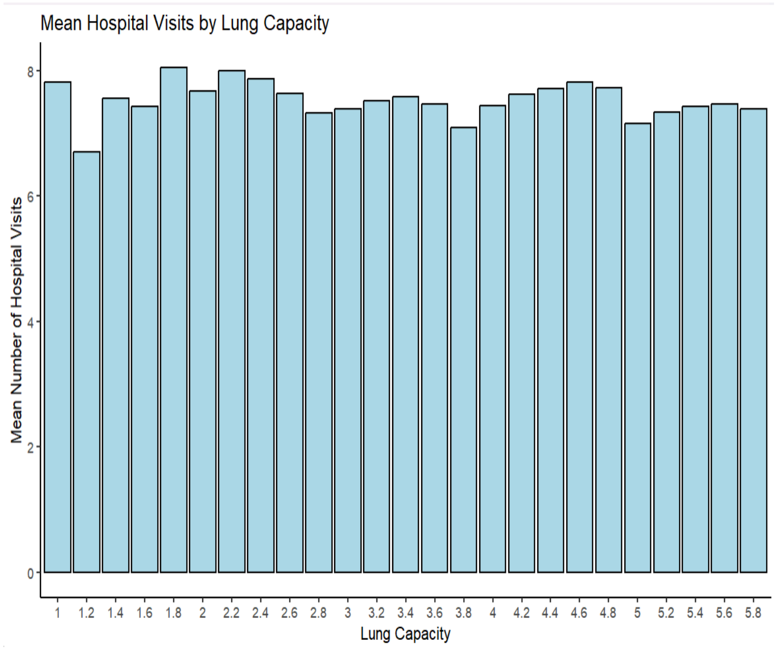

Figure 5. Mean Number of Hospital Visits by Lung Capacity.. The histogram displays the mean number of hospital visits for lung capacity. Lung capacity ranges from 1.00 to 6.00 liters, increasing in 0.2-liter intervals. The mean number of hospital visits remains relatively constant across all lung capacities. Even with the decrease in lung capacity, there is no change in the mean number of hospital visits.

A One-Way ANOVA was conducted to determine whether the lung capacity of individuals with lung disease affects the number of hospital visits. The test had a p-value of 0.477. This indicates that there is no statistically significant difference in the number of hospital visits for differing lung capacities. These findings suggest that lung capacity does not influence the number of hospital visits for an individual with lung disease.

Recovery Analysis

After cleaning (excluding missing or non-binary recovery entries), the dataset comprised 4345 patients with complete information on hospital visits and recovery status. Hospital visits ranged from 0 to 14 per patient (median = 8, IQR = 7), and overall, 51.02% of patients achieved full recovery. A bar plot of recovery rate by number of hospital visits (Fig. 6) shows that recovery proportions fluctuate around the overall mean, without a monotonic increase or decrease as visit count rises.

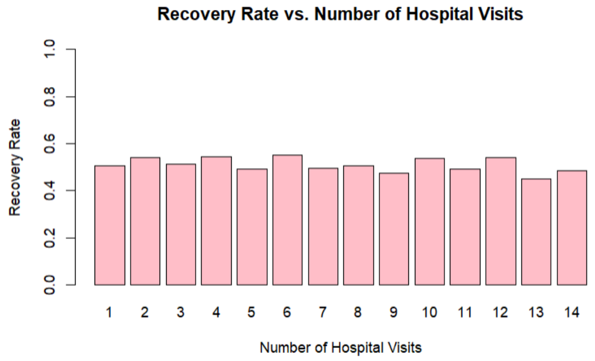

Figure 6. Bar plot of recovery rate by number of hospital visits. The plot shows how recovery rates vary across different visit counts. Recovery proportions stay close to the overall average and do not follow a clear upward or downward pattern as hospital visits increase.

A logistic regression model was fit to the number of hospital visits predicting recovery. The model coefficient for Hospital. Visits are x1 = -0.012. The impact is not statistically significant with a P value of 0.1069. Below is a graph of the model to visually show the lack of a significant impact.

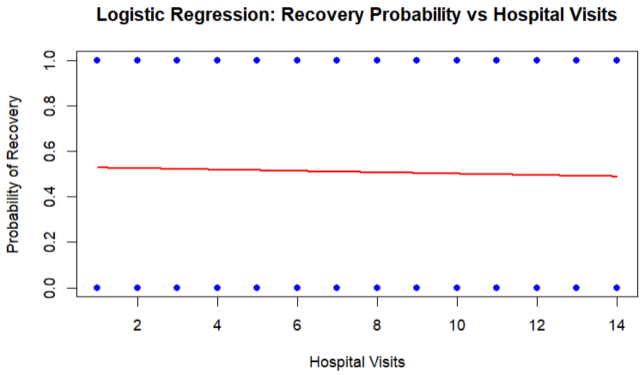

Figure 7. Logistic regression model predicting recovery probability by number of hospital visits. The red line shows the estimated probability of recovery across the range of hospital visits. The nearly flat line indicates that the number of hospital visits does not have a strong effect on recovery probability in this model.

To determine if hospital visits could have an impact in conjunction with disease type, lung capacity, and treatment type, the variables were added to the model via stepwise variable addition. The first step was to add a variable to each model, each with an interaction with Hospital. visits, and compare the AIC to the base model. Adding Lung Capacity produced the model with the lowest AIC. Subsequently, adding Disease Type reduced the AIC further. This was the final step, as adding another variable would increase the AIC, ending stepwise variable addition.

As shown in the final model, no variable has a single variable with a P value under 0.05, and therefore, we cannot reject the null hypothesis or prove that any of these variables have a statistically significant impact on the probability of recovery.

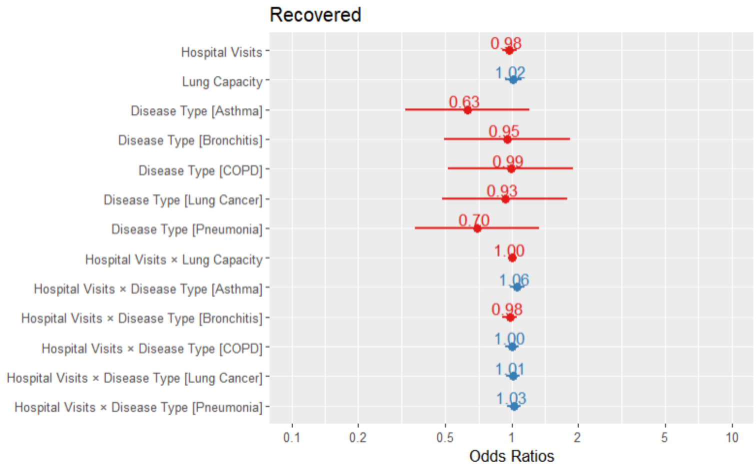

Figure 8. Odds ratios and 95% confidence intervals for predictors of recovery. Odds ratios less than 1 indicate a negative association with recovery. Notably, asthma and pneumonia show significantly reduced odds of recovery, as indicated by confidence intervals that do not cross 1. Interaction terms show minimal deviation from the null value, suggesting limited moderation effects.

To further attempt to predict the impact of these variables on recovery, stepwise variable elimination was used. First, the variable Hospital.Visits was removed along with any interactions with this variable. The remaining variables, Lung.Capacity + Disease.Type + Treatment.Type were removed one at a time, and the model with the lowest AIC was used for each step. None of the models obtained from this method have any statistically significant variables ($p < 0.05$).

Even after including additional predictors (Lung Capacity and Disease Type) and their interactions with Hospital. Visits, no individual coefficient meets the conventional significance level ($p < 0.05$). This means that with the data at hand, we cannot statistically reject the null hypothesis for any of these predictors. In other words, the analysis does not provide sufficient evidence to conclude that hospital visits, lung capacity, or disease type (or their interactions) have a statistically significant impact on the probability of recovery. Even when we expand the model to include interactions with lung capacity and disease type, none of these variables reaches statistical significance under conventional thresholds. Based on this analysis, we cannot reject the null hypothesis for these predictors.

Conclusion

This study explored four key questions under the broader aim of understanding what lifestyle or biological factors influence disease severity, healthcare utilization, or recovery outcomes among patients with lung disease. Specifically, we examined: (1) whether smoking status affects hospital visits or recovery, (2) whether age and gender influence the likelihood of developing specific types of lung disease, (3) whether lung capacity impacts the number of hospital visits, and (4) whether hospital visits—together with lung capacity, disease type, or treatment type—influence recovery outcomes. Across all four questions, we did not find strong statistical evidence supporting the expected relationships.

First, smoking status was not found to significantly influence either the number of hospital visits or recovery rates. Both ANOVA and Chi-square tests returned non-significant results, suggesting that in this dataset, smoking status alone does not account for variation in healthcare use or recovery outcomes. This contrasts with existing research showing a strong link between smoking and poorer health outcomes, which may indicate that other variables, such as severity or treatment quality, need to be considered.

Second, age and gender were not significant predictors of lung disease type in the multinomial logistic regression. While there were slight visual differences in disease distribution by age and gender, these patterns were not statistically significant. This could be because the dataset only includes individuals already diagnosed with a lung condition, limiting the ability to detect risk factors for disease development.

Third, lung capacity did not significantly influence the number of hospital visits. Although some variation was observed, a one-way ANOVA confirmed that these differences were not statistically meaningful. This finding suggests that lung capacity, by itself, may not drive hospital utilization once a lung disease is present.

Finally, a logistic regression model, with stepwise addition of lung capacity, disease type, and treatment type, was used to predict recovery. None of the variables or their interactions with hospital visits were statistically significant. This suggests that these clinical and demographic factors may not play a strong role in predicting recovery, at least not without considering additional variables like disease severity, adherence to treatment, or socioeconomic status.

These results go against common ideas about lung disease. We did not find a strong link between smoking and hospital visits or recovery. This may be because the dataset only shows whether someone smokes. It does not include details like how much or how long they smoked. We also found no clear effect of age or gender on disease type. One reason may be that everyone in the dataset already had a lung disease. The data cannot tell us who is more likely to develop these diseases. It only shows differences between types of illness. Lung capacity also showed no connection to hospital visits. This may be because hospital use depends on many things. People may visit the hospital due to habits, access to care, or insurance, not just lung function. The recovery model did not find any strong predictors either. This might be due to the way recovery was measured. The dataset only says “yes” or “no” for recovery. It does not tell us how long recovery took or how complete it was. The dataset has other limits. It does not show disease severity. It does not include whether patients followed treatment. It also does not account for other health problems. The data only covers one point in time. It does not follow patients over a longer period. Future research should collect more detailed information. It should track patients over time. This will help us learn how different factors affect lung disease and recovery. At the same time, the results remind us that lung disease is complex. Many different things can affect outcomes. These include personal habits, treatment, and access to care. Simple data may not be enough to capture this complexity. Researchers need better tools and more complete information. This will help improve decisions for people with lung disease.

Reference

American Lung Association. (n.d.). Estimated prevalence and incidence of lung disease. Retrieved April 13, 2025, from https://www.lung.org/research/trends-in-lung-disease/prevalence-incidence-lung-disease

Centers for Disease Control and Prevention. (2023). Chronic obstructive pulmonary disease (COPD). FastStats. Retrieved April 13, 2025, from https://www.cdc.gov/nchs/fastats/copd.htm

Centers for Disease Control and Prevention. (2023). Chronic obstructive pulmonary disease (COPD) and smoking. Retrieved April 13, 2025, from https://www.cdc.gov/tobacco/about/cigarettes-and-copd.html

Johns Hopkins Medicine. (n.d.). Smoking and respiratory diseases. Retrieved April 13, 2025, from https://www.hopkinsmedicine.org/health/conditions-and-diseases/smoking-and-respiratory-diseases

U.S. Department of Health and Human Services. (2010). How tobacco smoke causes disease: The biology and behavioral basis for smoking-attributable disease. Centers for Disease Control and Prevention. Retrieved April 13, 2025, from https://www.ncbi.nlm.nih.gov/books/NBK53021/

World Health Organization. (2023, November 15). Smoking is the leading cause of chronic obstructive pulmonary disease. Retrieved April 13, 2025, from https://www.who.int/news/item/15-11-2023-smoking-is-the-leading-cause-of-chronic-obstructive-pulmonary-disease