What is the Impact of Air Pollution on the Prevalence and Severity of Chronic Lung Disease?

Unurjargal Batsuuri

Introduction

Air pollution poses a significant health risk, particularly in densely populated urban areas like my hometown of Ulaanbaatar, Mongolia, known for high pollution levels. This concern has motivated my study on air pollution's impact on lung health.

The connection between air pollution and respiratory health has been substantiated by key studies such as those by Ko & Hui (2012), Duan, Hao, & Yang (2020), and Sunyer et al. (2000). These analyses illuminate the detrimental effects of pollutants on chronic obstructive pulmonary disease (COPD) and associated mortality risks, particularly with urban particulate matter. They establish the significant role of air pollution in respiratory disease exacerbation, yet they often employ binary or linear analytical methods that may not fully articulate the spectrum of disease severity or the variability in pollution exposure.

Our study advances this research by utilizing a multinomial logistic regression approach, which accommodates the gradations in lung disease severity and air pollution exposure. In contrast to prior work, this method considers the disease in three categories, low, medium, and high, and pollution across seven levels, providing a more sophisticated understanding of the relationship. It also integrates confounders such as genetics and lifestyle choices to offer a comprehensive view of lung health influencers.

The study stands to extend existing research, grounding its findings in both the global context and my personal hometown's environmental challenges, thereby highlighting the need for broader implications for public health and policy.

Methods

Data Description

Our study's dataset was extracted from a repository on lung cancer prediction, consisting of 1000 cases with complete data. The primary outcome variable was the level of chronic lung disease. Predictor variables included air pollution levels, demographic data such as age and gender, and health-related factors like smoking intensity and dust allergy levels.

Variable Selection and Statistical Analysis

A multinomial logistic regression model was employed to analyze the relationship between these predictors and the ordinal outcome of lung disease severity. To identify the most relevant predictors, we utilized both AIC and BIC in a stepwise selection process, supplemented by the LASSO technique. This combined approach aimed to optimize the model by including variables that contribute meaningfully to explaining the variation in lung disease severity while penalizing unnecessary complexity.

Model Diagnostics

Upon fitting model, diagnostics were done to ensure robust statistical inference. Correlation analysis between independent variables was the first step, to identify and address multicollinearity. Variables with high correlations without theoretical justification were considered for exclusion from the model. The next diagnostic step involved analyzing deviance residuals to detect outliers and influential data points, which could potentially skew our model's predictions. Residuals were plotted against predictors to assess their independence. Any identifiable patterns would prompt a re-evaluation of the data point's inclusion.

Model Validation

Model validation was assessed with 10-fold cross-validation to estimate predictive error, thus ensuring model consistency across different data subsets. To assess the model's calibration, we will construct plots that show Apparent and Bias-Corrected calibration lines against the Ideal line, which signifies perfect prediction. The convergence of these lines with the Ideal line would indicate precise calibration of probability predictions, a significant result for dependable classification of our outcome. Further, we implemented ROC curve analysis for discriminative validation. By considering each severity level as a binary outcome against the others, ROC curves

were constructed, and the corresponding AUC was calculated. An AUC value closer to 1 indicated excellent model performance, while a value near 0.5 suggested no discriminative ability beyond random chance.

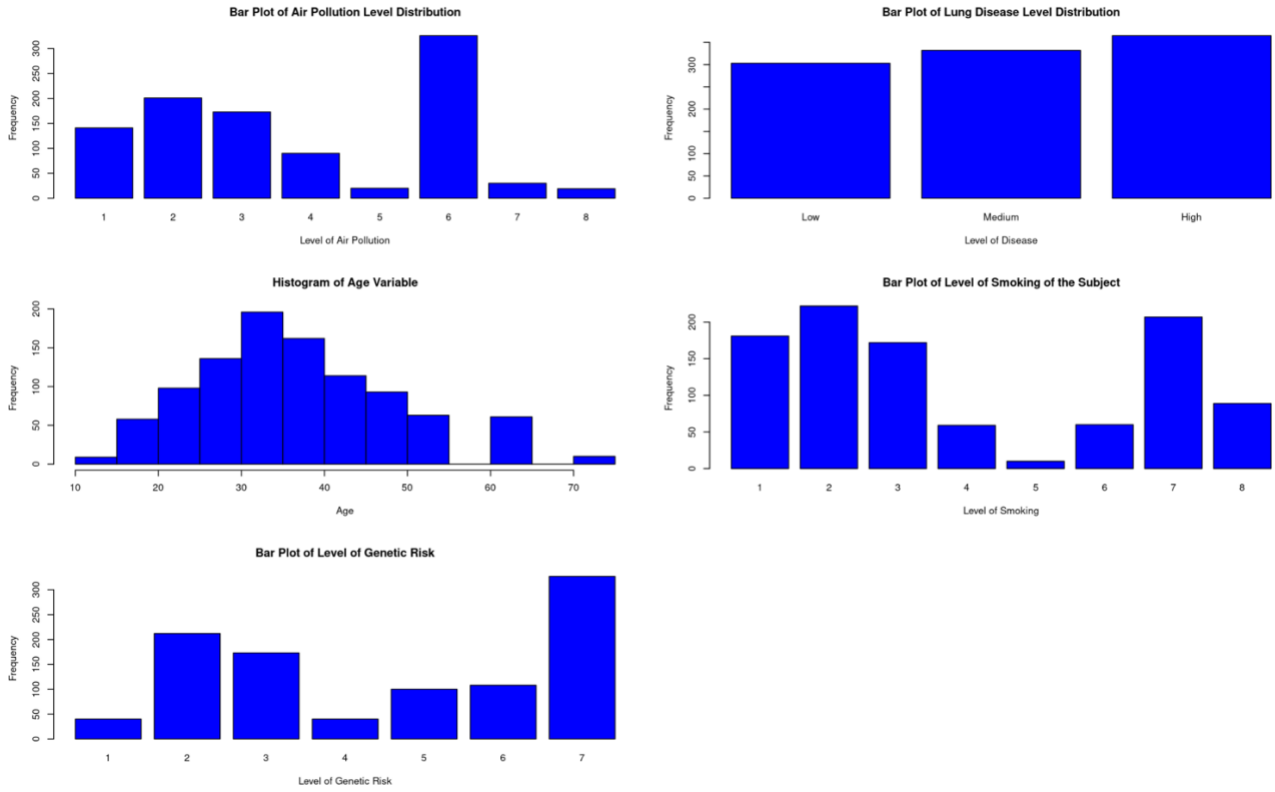
Each validation step was designed to confirm that our model is a reliable instrument for estimating probabilities across the spectrum of chronic lung disease severity, with adjustments made as needed to improve model accuracy and integrity.

Result

Description of the Data

Analysis of the dataset revealed skewed air pollution exposure, with the majority at higher levels, and a prevalence of ‘medium’ severity in chronic lung diseases. Age distribution the only numerical variable for our model follows an approximately normal distribution. Smoking habits were diverse, with notable proportions at specific exposure levels, and genetic risk factors were predominantly high. Plots 1 – 5 from Table 1 reveal these trends, with disproportionate sample sizes across categories that could potentially bias estimates and diminish statistical power for true association detection. Such imbalances were considered in subsequent modeling to ensure robust analysis.

Table 1



(Visual summaries of each variable)

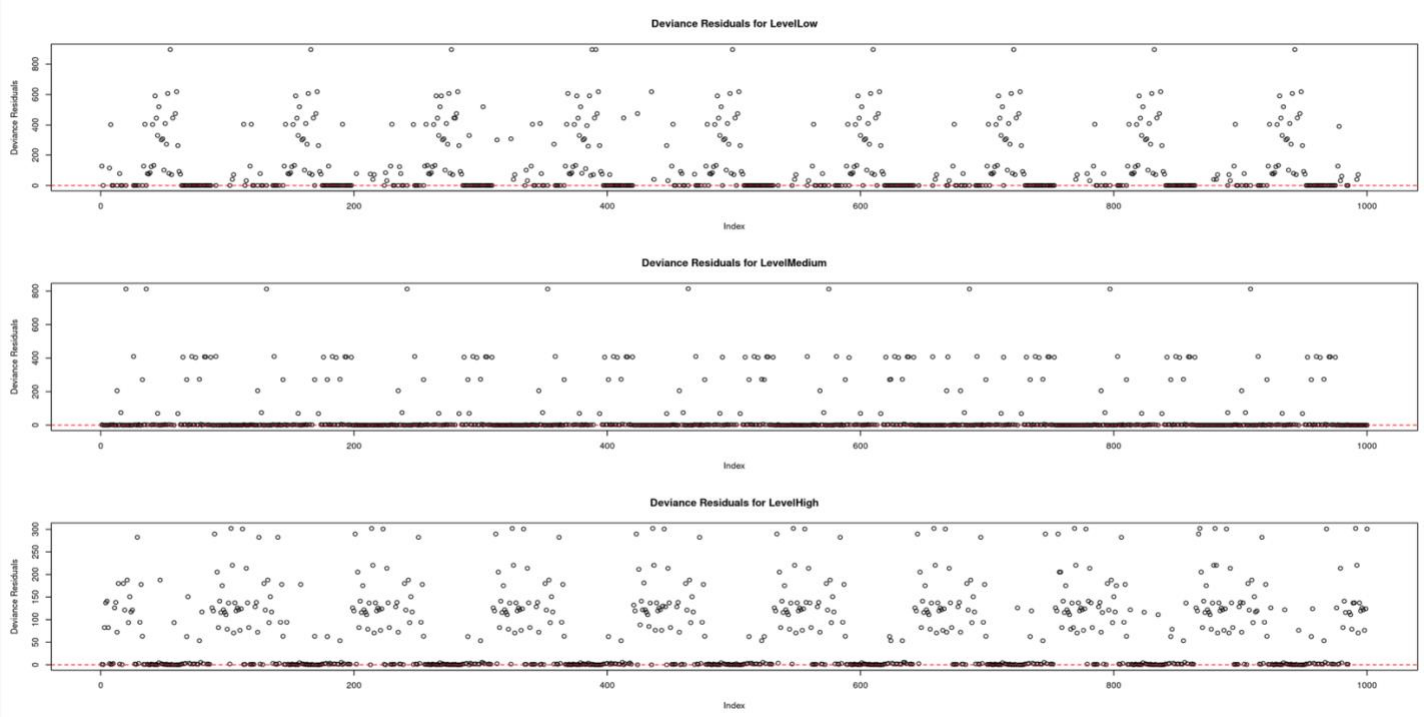
The Analysis Process and the Results

Our analytic process began with variable selection for our multinomial logistic regression model. The outcome variable was the level of lung disease, with age, smoking, genetic risk, and occupational hazard as potential predictors. While the AIC and BIC selection processes did not initially identify air pollution as a significant variable, LASSO analysis indicated its importance for higher levels of lung disease severity. Consequently, air pollution was manually incorporated into the final model. Occupational hazard was excluded due to its high correlation with genetic risk, evidenced in the appendix Figure 1, and the lack of substantial literature to justify its inclusion.

Continuing, deviance residuals, illustrated in Figure 2, showed a random distribution around zero for ‘Low’ and ‘High’ severity categories, indicating a good fit for these data points. A few observations had large residuals, suggesting potential outliers. The ‘Medium’ category presented

a tighter cluster of residuals, suggesting consistent model performance for this group or a potential need for category-specific model adjustments. No removal of data points was executed to preserve the dataset's variance and enhance model robustness. Outliers were kept under advisement for further examination to confirm their validity and ensure comprehensive representation in the model.

Figure 3. Deviance residuals for all the levels of the outcome variable



Validation of the Final Model

The final model achieved an AUC greater than 90% as per Figure 3, suggesting an excellent capacity to differentiate between the disease severity levels. This robust performance is critical for a model that is expected to accurately predict multiple categories within a multinomial context.

Calibration evaluation, shown in Figure 4, indicated a well-calibrated model, as apparent, bias-corrected, and ideal lines were closely aligned. This confluence, especially near the ideal 45-degree line, affirms the model's predictions are close with actual outcomes, enhancing confidence in the model's applicability. These validation findings illustrate a model that not only discriminates

between the lung disease severity levels effectively but also calibrates the predicted probabilities with observed outcomes accurately.

Figure 3. ROC plots for the Levels of Lung Disease Severity

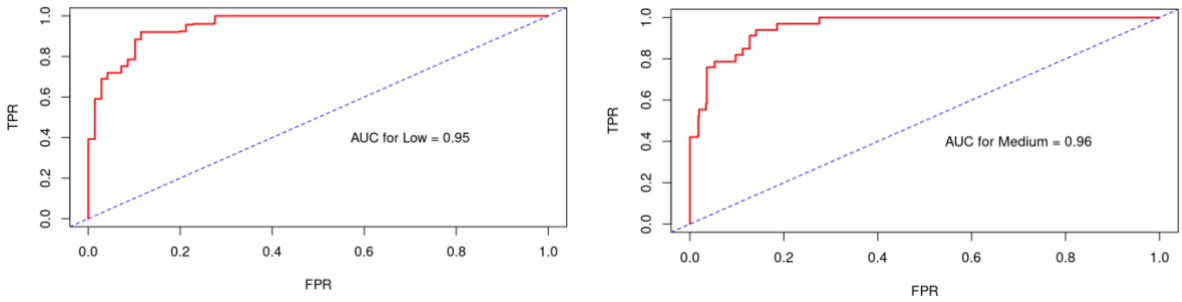
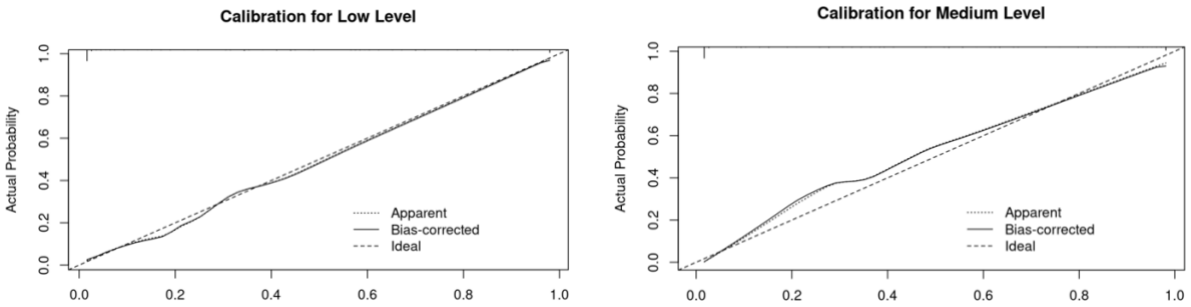


Figure 4. Calibration Plots for the Levels of Lung Disease Severity



Discussion

Our multinomial logistic regression analysis revealed a complex relationship between air pollution levels, demographic factors, lifestyle choices, and the prevalence and severity of chronic lung disease. We found that more pollution is often linked to worse lung disease. Particularly, higher coefficients for pollutants at levels four through seven were significantly, with p-values around 0, associated with severe lung disease. The high coefficient reflects the strength of the association between this specific level of air pollution and the outcome of interest, lung disease severity

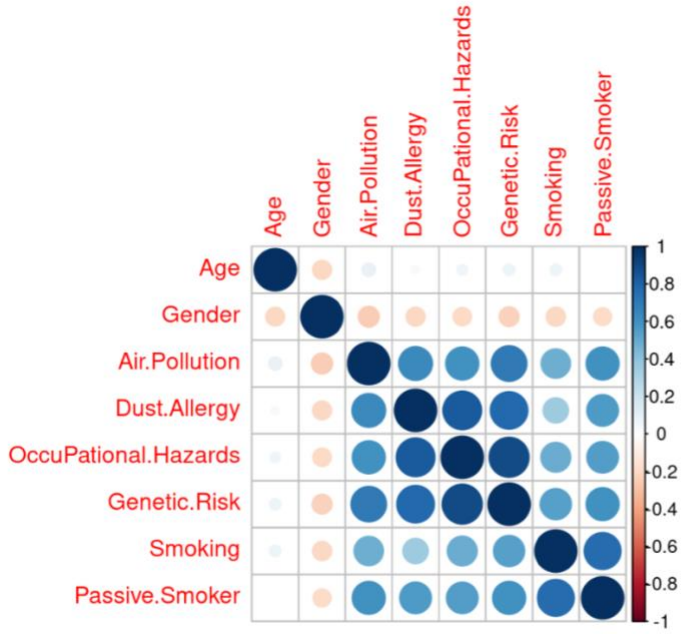
level.This correlation is a critical finding, where greater exposure leads to worse health outcomes. The model also Identified genetic risk as a significant factor affecting lung disease severity.

This aspect points to the additive effects of genetic susceptibility and environmental factors, indicating the need for targeted public health measures. However, the cross-sectional design of our study constrains us from establishing causality. Longitudinal research is essential for understanding the dynamics of pollution exposure and its health impacts over time. Our analysis used a multinomial rather than an ordinal regression, influenced by course material limitations, which might not fully capture the ordered nature of lung disease severity. One particular challenge was the inability to generate a calibration plot for the high severity level of lung disease due to issues with fitting the model using **lrm.fit** in R. This technical difficulty limits our capacity to fully validate the model's performance for this group, which is a significant concern, as it impedes our ability to assess the model's accuracy in predicting the most severe outcomes. The robust association between pollutant levels and disease severity aligns with the goal of our research to prove the links between environmental factors and health.

In conclusion, our model, while indicative of the harmful effects of air pollution, prompts further research to explore this complex relationship and reinforce the basis for preventive strategies in environmental health.

Appendix

Figure 1. Correlation plot of the variables



Note: On the right we can see the levels of correlation based on colour, going from -1 to 1

References

Duan, R.-R., Hao, K., & Yang, T. (2020). Air pollution and chronic obstructive pulmonary disease. Chronic Diseases and Translational Medicine, 6(4), 260–269. https://doi.org/10.1016/j.cdtm.2020.05.004

Sunyer, J., Schwartz, J., Tobías, A., Macfarlane, D., Garcia, J., & Antó, J. M. (2000). Patients with Chronic Obstructive Pulmonary Disease Are at Increased Risk of Death Associated with Urban Particle Air Pollution: A Case-Crossover
Analysis. American Journal of Epidemiology, 151(1), 50–56. https://doi.org/10.1093/oxfordjournals.aje.a010121

KO, F. W. S., & HUI, D. S. C. (2012). Air pollution and chronic obstructive pulmonary disease. Respirology (Carlton,Vic.), 17(3), 395–401. https://onlinelibrary.wiley.com/doi/10.1111/j.1440-1843.2011.02112.x

TheDevastator. (n.d.). Lung Cancer Prediction: A new link [Data set]. Kaggle. Retrieved [2023], from https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link/data

Baker, A. (2018, March 1). This is the world's most polluted capital, where babies wear face masks to daycare. Time https://time.com/longform/ulan-bator-mongolia-most-polluted-capital/