Data description

1.0 Dataset

This study utilizes the IPUMS NHIS (National Health Interview Survey) dataset, which provides comprehensive health data typical of the U.S. population. The dataset is suitable for analysing the relationship between smoking habits and lung cancer risk, including detailed self-reported smoking histories, demographic information, and health statuses.

The study focuses on current smokers, omitting former smokers to reduce the impact of smoking cessation on smoking duration and lung cancer risk. Data from three separate time periods (2000, 2005, and 2010) were collected, ensuring a sufficient sample size and temporal diversity for analysis.

2.0 Dependent variable

The dependent variable in this study is lung cancer diagnosis (CNLUNG), represented as a binary variable indicating the pres**ence of a lung cancer diagnostic (1 = Yes, 0 = No). CNLUNG in the original dataset identifies adult samples who have been diagnosed with lung cancer at any point. After excluding invalid and unanswered samples, the remaining s**amples were categorized into two categories, 0 and 1, based on their responses.

3.0 Independent variables

The model identifies smoke duration as the independent variable. Smoking duration denotes the cumulative years from the initiation of smoking to the diagnosis of lung cancer. The duration of smoking may elevate the risk of lung cancer, thus serving as a variable to assess the influence of smoking length on lung cancer incidence. To establish the variable of smoke duration, two metrics—age (AGE) and the age at which smoking was first engaged in regularly (SMOKAGEREG)—are utilized for the calculation of smoke duration. Consequently, as previously stated, all statistics pertain to present smokers, and smoke duration is calculated as AGE minus SMOKAGEREG.

4.0 Control variables

To mitigate the influence of confounding variables on the outcomes, we incorporated the subsequent control variables into the model, as informed by prior research. This study incorporated three control variables into the model: gender, income, and family cancer history.

Gender (SEX) is a binary variable that differentiates between male and female respondents. In the IPUMS NHIS dataset, sex is self-reported and categorized as "male" (Male=1) or "female" (Female=2). The sample responses comprised individuals who declined to disclose information, those who expressed uncertainty, and those who were unaware, all of which were excluded from the

dataset due to their inapplicability to the data analysis of this study. Sex was incorporated as a control variable due to prior research indicating that biological variations and sex-specific behaviors influence smoking patterns and vulnerability to lung cancer. By adjusting for sex, we sought to isolate the impact of smoking habit on lung cancer risk across various sex groups.

Income level is a categorical variable denoting an individual's total personal earnings from the preceding calendar year. The variable EARNIMP1 encompasses both reported and imputed values to rectify missing data, hence providing comprehensiveness in income-related analysis. This imputation adheres to the approach specified in the National Health Interview Survey (NHIS) imputation documentation, which indicates that non-response rates for income-related inquiries are generally elevated. The imputed income variables in the IPUMS NHIS dataset have been integrated with other survey data from each survey year, facilitating consistent analysis. The dataset categorizes EARNIMP1 as a numeric variable, with 31 specified income levels ranging from $01 to above $115,000. Income level is a vital control variable as it affects both smoking habit and health effects. Previous research indicates that individuals in lower-income categories may exhibit increased smoking rates and restricted access to healthcare services, potentially affecting lung cancer detection and treatment results. This study seeks to regulate income levels to mitigate the socioeconomic factors that may obscure the correlation between smoking duration and lung cancer risk.

The family history of lung cancer variable is formulated as a binary indicator to reflect hereditary susceptibility to lung cancer. This variable is extracted from six associated variables in the IPUMS NHIS dataset, indicating whether a respondent's biological family, including mother, father, full siblings, and children, had received a lung cancer diagnosis. The variables BMLGCAN, BFLGCAN, BDLGCAN, BSLGCAN, FBLGCAN, and FSLGCAN denote if the respondent's biological mother, father, daughter, son, full brother, or full sister, respectively, had lung cancer. The family_lung_cancer variable is assigned a value of 1 if any of the six specified variables are entered as “Mentioned” (Value = 2), signifying that at least one direct family member has been diagnosed with lung cancer. Conversely, it is designated as 0 if all six factors are documented as “Not mentioned” (Value = 1), signifying no reported familial history of lung cancer. Responses classified as “Unknown,” “Refused,” or “Not ascertained” (Values = 7, 8, 9) are omitted from the analysis to ensure data integrity. Incorporating familial lung cancer as a control variable facilitates the adjustment of putative genetic factors on lung cancer risk, which is essential for isolating the impact of smoking duration. Previous studies indicate that individuals with a familial predisposition to lung cancer possess an elevated baseline risk attributable to inherited genetic characteristics, common environmental exposures, or analogous lifestyle practices. This study seeks to compensate

for confounding factors by incorporating this variable, so facilitating a more precise evaluation of the relationship between smoking behavior and lung cancer risk.

Description of How to Run the Code

To replicate the analysis in this report, follow these steps:

1. Import Dataset: Load the dataset nhis_00014.csv into R using read.csv().
2. Data Cleaning and Preprocessing
    - Follow the sequence in the R Markdown file to clean and preprocess the data.
    - Process key variables in the following order:
        - Smoking Duration
        - Family Lung Cancer History
        - CNLUNG (Lung Cancer Diagnosis)
3. Descriptive Statistics
    - Compute summary statistics (mean, standard deviation, frequency tables) for key variables using functions like summary() and table().
4. Logistic Regression Model
    - Fit a logistic regression model using the glm() function with binomial(link="logit").
    - Extract regression output and interpret coefficient estimates.
    - Assess multicollinearity using Variance Inflation Factor (VIF).
    - Compute heteroskedasticity-robust standard errors to account for potential violations of homoscedasticity.
5. Robustness Check
    - Estimate multiple models by adding and removing variables to test the stability of coefficient estimates.
    - Compare model outputs (coefficient estimates, p-values, AIC, and deviance) to evaluate the robustness of findings.
6. Heterogeneity Analysis
    - Fit separate logistic regression models for males and females to assess gender differences in smoking-related lung cancer risk.
    - Compare coefficient estimates and p-values to analyze heterogeneity.
    - Introduce an interaction term (smoking duration × sex) in the logistic model to formally test for gender-based differences in smoking effects.