Dean&Francis

ISSN 2959-6157

# Study of Lifestyle Habits Affecting Lung Cancer

**Yingchi Zeng**[1, *]

[1]Department of Bioengineering, Shanghai University, Shanghai, 200444, China

*Corresponding author: zyc101123@shu.edu.cn

**Abstract:**

Lung cancer is one of the most lethal forms of cancer and the primary cause of cancer-related mortality on a global scale. Given the rising incidence of lung cancer on an annual basis, it is of paramount importance to investigate further the potential risk factors associated with this disease in order to develop effective, personalized prevention strategies. In this study, a binary logistic regression model is used to predict the risk of lung cancer in patients by analyzing demographic and medical data. The dataset consisted of 300 study participants and 15 variables, such as smoking and gender. The dependent variable is whether the patient has lung cancer. In this study, smoking, peer pressure, chronic disease, fatigue, allergy, coughing, and swallowing difficulty have a significant positive effect on lung cancer. Furthermore, the overall predictive accuracy of the study model is 94.33%. Therefore, the predictive results of the logistic regression model are acceptable. In order to better predict lung cancer occurrence, more comprehensive clinical data and more advanced analysis techniques are needed, and more influencing factors need to be taken into account. The model lays the foundation for the prediction of lung cancer.

**Keywords:** Lung cancer; logistic regression; prediction model.

## 1. Introduction

Lung cancer represents a significant global health concern. It is estimated that approximately 2.1 million individuals are diagnosed with lung cancer annually, with 1.8 million deaths attributed to this disease [1]. The incidence of lung cancer continues to increase globally, although in some Western countries, the incidence in men is declining. Approximately 85 percent of lung cancers are non-small cell lung cancers (NSCLCs), while the remaining 15 percent are small cell lung cancers (SCLCs). Therefore, an in-depth investigation of the relevant factors affecting the occurrence of lung cancer not only helps scientists to understand lung cancer pathological mechanisms more comprehensively, but also is of great significance for the development of effective preventive measures, early diagnosis and personalized treatment plans.

Currently, the main factors affecting the development of cancer include tobacco use, alcohol intake, unhealthy diet, and lack of physical activity [2-4]. Furthermore, research is being conducted into the effects of air pollution and chronic infectious factors. It has been demonstrated that smoking represents the primary cause of lung cancer [5]. Irrespective of the specific type of lung cancer and the stage of the disease, the continuous consumption of tobacco products has a detrimental impact on its progression and the efficacy of any subsequent treatment. In addition, exposure to second-hand smoke has been observed to elevate the risk of developing lung cancer. Consequently, cessation of smoking and avoidance of second-hand smoke exposure represent a crucial strategy for the prevention of lung cancer [6]. What's more, environmental pollution represents a further significant factor affecting the development of lung cancer. The occurrence of lung cancer is closely related to the presence of pollutants such as particulate matter, sulfur dioxide, and nitrogen oxides in the atmosphere. Those who are exposed to these pollutants over an extended period of time are at a markedly elevated risk of developing lung cancer. Additionally, genetic factors have been demonstrated to exert a contributory influence on the development of lung cancer [7]. These mutations or genetic variants may contribute to an individual's susceptibility to this disease. Last but not least, occupational exposure represents a significant contributing factor in the development of lung cancer [8]. The presence of certain chemicals and radioactive substances in certain occupational environments has been linked to an increased risk of developing lung cancer. Therefore, strengthening occupational health surveillance and protective measures is one of the important means to prevent lung cancer.

In addition, choosing the right data model is crucial for analyzing the factors affecting lung cancer development. First of all, one needs to understand the pathogenesis of lung cancer and the factors affecting it. Lung cancer is a complex disease, and its pathogenesis involves multiple factors such as genetics, environment, and lifestyle. Therefore, multivariate data models are commonly used to analyze the influence of these factors on lung cancer. In terms of data regression model selection, Population sample of 3,177 cancer-free adults from the Baltimore Epidemiologic Catchment Area Study (BECAS) was selected by Alden et al. and followed for up to 24 years. Then the Cox proportional risk model was used to estimate the relative risk of overall and subtype-specific cancers in patients with a history of depression. The final conclusion was that there is a correlation between depression and hormone-mediated cancer [9]. You et al. employed polygenic risk score (PRS) and genome-wide interaction analysis (GWIA) to evaluate the multiplicative interactions between BMI trajectories and genetic variants associated with the risk of NSCLC. Their findings revealed a correlation between BMI trajectories, genetic factors, and the risk of developing NSCLC [10]. However, these studies only performed correlation analyses for a single factor or several factors, which were not comprehensive, so this study chose to explore the effects of these factors on lung carcinogenesis using a binary logistic regression model.

# 2. Methods

## 2.1 Data source

The data presented in this literature review was sourced from the Kaggle website and collated by Shreyas Paraj-Patil's team. It was last updated in February 2024 and encompassed the responses of 300 individuals. It gives examples of various symptoms of lung cancer which have been carefully curated to cover a wide range of symptoms and ensures that the models generated are versatile and accurate.

## 2.2 Variable Selection

The sample size of this dataset is 300, with 156 men and 144 women. The dataset consists of fifteen variables as shown in Table 1, which are defined as $X_1$ to $X_{15}$ for ease of writing and where 0 means normal and 1 means suffering from the disease.

**Table 1. Variable description**

| Term | Type | $X_n$ | Description |
|---|---|---|---|
| Gender | Categorical | $X_1$ | 0-Female , 1-male |
| Age | Numeric | $X_2$ | 21 - 87 |

| | | | |
|---|---|---|---|
| Smoking | Categorical | $X_3$ | 0-NO, 1- YES |
| Yellow fingers | Categorical | $X_4$ | |
| Anxiety | Categorical | $X_5$ | |
| Peer Pressure | Categorical | $X_6$ | |
| Chronic Diseases | Categorical | $X_7$ | |
| Fatigue | Categorical | $X_8$ | |
| Allergies | Categorical | $X_9$ | |
| Wheezing | Categorical | $X_{10}$ | |
| Alcohol Consuming | Categorical | $X_{11}$ | |
| Coughing | Categorical | $X_{12}$ | |
| Shortness of Breath | Categorical | $X_{13}$ | |
| Swallowing Difficulty | Categorical | $X_{14}$ | |
| Chest Pain | Categorical | $X_{15}$ | |

### 2.3 Method Introduction

In this paper, Binary Logistic Regression modeling is used, with 15 factors as independent variables (X) and the presence of lung cancer as dependent variable (Y). Here 0 means no and 1 means yes. The logistic regression model can be represented as:

$$ln\left(\frac{p}{1-p}\right)=g\left(x\right)=\beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_nX_n \quad (1)$$

$p$ denotes the probability that the dependent variable $Y$ is 1 given the independent variable $X$ ; $1-p$ represents the probability that Y does not occur under condition X; $\beta_0$, $\beta_1$, $\beta_2$, ..., $\beta_n$ are model parameters to be estimated from training data; $X_1, X_2$, ..., $X_n$ are independent variables.

## 3. Results and Discussion

### 3.1 Descriptive Analysis

From table 2, it can be seen that the proportion of males is more than females, 52.00 percent and 48.00 percent respectively. From the fatigue distribution, the majority of the samples are ‘1.0’, i.e., there are 201 persons suffering from fatigue, which is 67.00 percent. The percentage of those without fatigue is 33.00%. Also, the number of people suffering from shortness of breath is more than half of the sample with a percentage of 64.00%. The ratio of diseased/normal numbers for the rest of the variables is close to 1:1. It is noteworthy that the present study investigated the number of people who suffer from lung cancer is nearly 90%, which is the majority of the population with 261 people.

Figure 1 shows a histogram of the age distribution of the sample. The data shows that the age range of the sample is between 21-87, with the majority of the sample concentrate in the 60-70 age range.

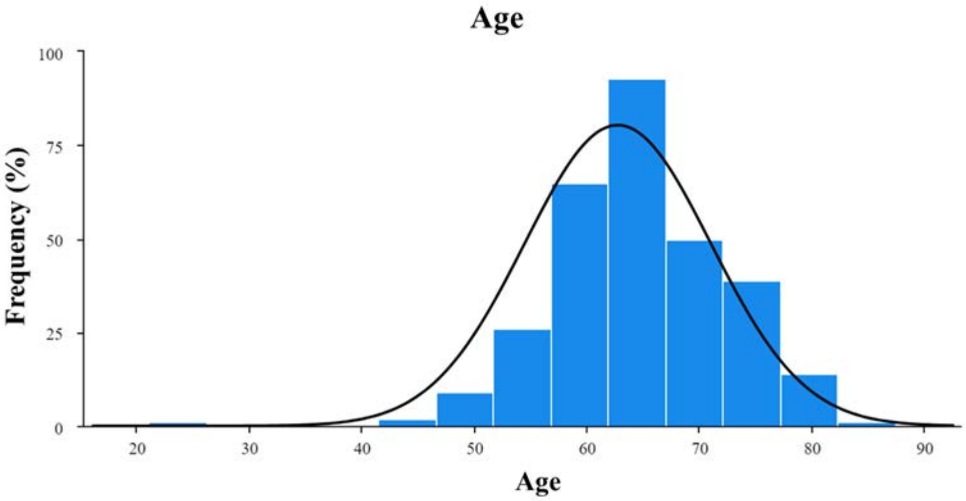

**Fig. 1 Histogram of age distribution**

**Table 2. Frequency analysis results**

| Name | Value | Frequency | Percentage (%) | Cumulative percentage (%) |
|---|---|---|---|---|
| $X_1$ | 0.0 | 144 | 48.00 | 48.00 |
| | 1.0 | 156 | 52.00 | 100.00 |
| $X_3$ | 0.0 | 131 | 43.67 | 43.67 |
| | 1.0 | 169 | 56.33 | 100.00 |
| $X_4$ | 0.0 | 128 | 42.67 | 42.67 |
| | 1.0 | 172 | 57.33 | 100.00 |
| $X_5$ | 0.0 | 150 | 50.00 | 50.00 |
| | 1.0 | 150 | 50.00 | 100.00 |
| $X_6$ | 0.0 | 150 | 50.00 | 50.00 |
| | 1.0 | 150 | 50.00 | 100.00 |
| $X_7$ | 0.0 | 147 | 49.00 | 49.00 |
| | 1.0 | 153 | 51.00 | 100.00 |
| $X_8$ | 0.0 | 99 | 33.00 | 33.00 |
| | 1.0 | 201 | 67.00 | 100.00 |
| $X_9$ | 0.0 | 135 | 45.00 | 45.00 |
| | 1.0 | 165 | 55.00 | 100.00 |
| $X_{10}$ | 0.0 | 134 | 44.67 | 44.67 |
| | 1.0 | 166 | 55.33 | 100.00 |
| $X_{11}$ | 0.0 | 135 | 45.00 | 45.00 |
| | 1.0 | 165 | 55.00 | 100.00 |
| $X_{12}$ | 0.0 | 128 | 42.67 | 42.67 |
| | 1.0 | 172 | 57.33 | 100.00 |
| $X_{13}$ | 0.0 | 108 | 36.00 | 36.00 |
| | 1.0 | 192 | 64.00 | 100.00 |
| $X_{14}$ | 0.0 | 160 | 53.33 | 53.33 |
| | 1.0 | 140 | 46.67 | 100.00 |

| | | | | |
|---|---|---|---|---|
| $X_{15}$ | 0.0 | 134 | 44.67 | 44.67 |
| | 1.0 | 166 | 55.33 | 100.00 |
| Lung Cancer | 0.0 | 39 | 13.00 | 13.00 |
| | 1.0 | 261 | 87.00 | 100.00 |
| Total | | 300 | 100.0 | 100.0 |

### 3.2 Logistic Regression Results

In order to conduct a binary logit regression analysis, it is necessary to take the aforementioned variables and use them as independent variables, while utilising lung cancer as the dependent variable. As evidenced in Table 3, a total of 300 samples are included in the analysis, with no instances of missing data.

**Table 3. Basic Summary of Binary Logit Regression Analysis**

| Name | Options | Frequency | Percentage |
|---|---|---|---|
| Lung Cancer | 0 | 39 | 13.00% |
| | 1 | 261 | 87.00% |
| | Total | 300 | 100.0% |
| Summary | Valid | 300 | 100.00% |
| | Missing | 0 | 0.00% |
| | Total | 300 | 100.0% |

The model validity is then analysed using the Model Likelihood Ratio test and it is found that with a p-value of less than 0.05 the model is valid and can be used in subsequent binary logistic regressions to determine the extent to which each variable influenced the dependent variable. In addition, the AIC and BIC values can be used for comparison in multiple analyses (table 4).

**Table 4. Results of likelihood ratio test for binary logit regression model**

| Model | -2 $x$ log likelihood | $\chi^2$ | d$f$ | p | AIC | BIC |
|---|---|---|---|---|---|---|
| Intercept only | 231.832 | | | | | |
| Final model | 91.515 | 140.317 | 15 | 0.000 | 123.515 | 182.775 |

Summarizing table 5, it is clear that $X_3, X_6, X_7, X_8, X_9, X_{12}, X_{14}$ have a significant positive influence relationship on lung cancer with dominance ratios (OR values) of 5.920, 5.655, 24.236, 21.428, 4.954, 26.825, 22.257, but $X_1, X_2, X_4, X_5, X_{10}, X_{11}, X_{13}, X_{15}$ don't have an impact relationship with lung cancer.

Thus, the model equation can be obtained as:

$$ln\left(\frac{p}{1-p}\right) = -8.329 - 0.536X_1 + 0.022X_2 + \ldots + 0.544X_{15} \tag{2}$$

**Table 5. Summary of results from binary logit regression analyses**

| Term | Marginal effect (dy/dx) | Standard error | z-value | Wald $\chi^2$ | p-value | OR-value | 95% CI |
|---|---|---|---|---|---|---|---|
| $X_1$ | -0.536 | 0.709 | -0.755 | 0.571 | 0.450 | 0.585 | 0.146 ~ 2.350 |
| $X_2$ | 0.022 | 0.034 | 0.641 | 0.411 | 0.521 | 1.022 | 0.956 ~ 1.092 |
| $X_3$ | 1.778 | 0.704 | 2.525 | 6.375 | 0.012* | 5.920 | 1.489 ~ 23.540 |
| $X_4$ | 1.384 | 0.744 | 1.859 | 3.455 | 0.063 | 3.989 | 0.928 ~ 17.158 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $X_{5}$ | 0.895 | 0.813 | 1.100 | 1.210 | 0.271 | 2.447 | 0.497 ~ 12.047 |
| $X_{6}$ | 1.732 | 0.662 | 2.617 | 6.851 | 0.009** | 5.655 | 1.545 ~ 20.692 |
| $X_{7}$ | 3.188 | 0.885 | 3.601 | 12.970 | 0.000** | 24.236 | 4.276 ~ 137.378 |
| $X_{8}$ | 3.065 | 0.824 | 3.717 | 13.819 | 0.000** | 21.428 | 4.258 ~ 107.834 |
| $X_{9}$ | 1.600 | 0.774 | 2.069 | 4.280 | 0.039* | 4.954 | 1.088 ~ 22.566 |
| $X_{10}$ | 0.968 | 0.831 | 1.164 | 1.355 | 0.244 | 2.632 | 0.516 ~ 13.424 |
| $X_{11}$ | 1.408 | 0.796 | 1.770 | 3.132 | 0.077 | 4.088 | 0.859 ~ 19.442 |
| $X_{12}$ | 3.289 | 1.067 | 3.082 | 9.499 | 0.002** | 26.825 | 3.312 ~ 217.259 |
| $X_{13}$ | -0.695 | 0.759 | -0.915 | 0.837 | 0.360 | 0.499 | 0.113 ~ 2.211 |
| $X_{14}$ | 3.103 | 1.129 | 2.748 | 7.551 | 0.006** | 22.257 | 2.434 ~ 203.477 |
| $X_{15}$ | 0.544 | 0.690 | 0.788 | 0.621 | 0.431 | 1.722 | 0.446 ~ 6.656 |
| Intercept | -8.329 | 2.501 | -3.330 | 11.089 | 0.001 | 0.000 | 0.000 ~ 0.032 |

Notes: Dependent variable = LUNG CANCER
McFadden $R^2$ = 0.605
Cox & Snell $R^2$ = 0.374

Nagelkerke $R^2$ = 0.694
* p<0.05 ** p<0.01 z-values in parentheses

**Table 6. Summary of Binary Logit Regression Prediction Accuracy**

| | | Prediction value | | Prediction accuracy | Prediction error rate |
|---|---|---|---|---|---|
| | | 0 | 1 | | |
| Real value | 0 | 29 | 10 | 74.36% | 25.64% |
| | 1 | 7 | 254 | 97.32% | 2.68% |
| Summary | | | | 94.33% | 5.67% |

The model prediction accuracy is used to judge the quality of model fitting. As can be seen from Table 6, the overall prediction accuracy of the research model is 94.33% and the model fit is relatively high, indicating that the binary logistic regression model constructed in the study can predict the occurrence of lung cancer more accurately.

**Table 7. Hosmer-Lemeshow goodness-of-fit test**

| $\chi^2$ | d*f* | p-value |
|---|---|---|
| 1.889 | 8 | 0.984 |

The fitted values of the model are found to be consistent with the observed values (table 7). The p-value is greater than 0.05 ($\chi^2$=1.889, p=0.984>0.05), indicating that the original hypothesis is accepted. This result demonstrates that the model passes the HL test and is well fitted, thereby providing evidence that the model can be used to predict the probability of lung cancer.

## 4. Conclusion

This study shows that seven variables (smoking, peer pressure, chronic disease, fatigue, allergy, coughing, swallowing difficulty) have a significant positive effect relationship on lung cancer, which is more likely to affect lung carcinogenesis, with an overall prediction accuracy is 94.33 percent. Data patterns are analyzed to identify key factors that contribute to an individual's susceptibility to heart disease. This will contribute to better preventive health measures for lung cancer patients in general or for specific high-risk groups (e.g., specific age or gender). In addition, the present research study is not comprehensive in the selection of independent variables. There are still other influencing factors, such as genes, obesity, air quality, etc. Therefore, the influence and interaction of environmental and genetic factors can be considered in future

ISSN 2959-6157

studies to seek to establish a more complete prediction model.

## References

[1] Bray F, Ferlay J, Soerjomataram I, et al. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin, 2018, 68(6): 394-424.

[2] Sun Q, Xie W, Wang Y, et al. Alcohol consumption by beverage type and risk of breast cancer: A dose-response meta-analysis of prospective cohort studies. Alcohol, 2020, 55(3): 246-253.

[3] Tran L, Bobe G, Arani G, et al. Diet and pparg2 pro12ala polymorphism interactions in relation to cancer risk: A systematic review. Nutrients, 2021, 13(1).

[4] Drope Jeffrey, Schluger Neil W. The tobacco atlas. Lancet Oncology, 2018.

[5] Yang Y, Cheng C, He B, et al. Cigarette smoking, by accelerating the cell cycle, promotes the progression of non-small cell lung cancer. Journal of Hazardous Materials, 2023, 455: 131556.

[6] Yorifuji T, Kashima S. Air pollution: Another cause of lung cancer. Lancet Oncology, 2013, 14(9): 788-789.

[7] Sampson J N, et al. Analysis of heritability and shared heritability based on genome-wide association studies for thirteen cancer types. J Natl Cancer Inst, 2015, 107(12): 279.

[8] Chandwani R, Brokamp C, Salfity H, et al. Impact of environmental exposures on lung cancer in patients who never smoked. World Journal of Surgery, 2023, 47(10): 2578-2586.

[9] Gross A L, Gallo J J, Eaton W. Depression and cancer risk: 24 years of follow-up of the baltimore epidemiologic catchment area sample. Cancer Causes Control, 2010, 21(2): 191-199.

[10] You D, et al. Associations of genetic risk, bmi trajectories, and the risk of non-small cell lung cancer: A population-based cohort study. BMC Medicine, 2022, 20(1): 203.