Research on Predicting Lung Cancer with Clinical Variables

Haotian Zhang *

School of Brunel London School, North China University of Technology, Beijing, 100000, China

* Corresponding Author Email: 23190010107@mail.ncut.edu.cn

Abstract. At present, research on lung cancer prediction mainly focuses on accurate prediction of lung cancer, and there is little research on predicting the risk of lung cancer based on clinical factors. In this study, a binary logistic regression model was used to process a dataset of lung cancer-related factors from Kaggle, which was uploaded to the website in 2022 and contains 309 samples and 15 influencing factors. Based on a logistic regression of 15 factors, it was found that only Smoking, Yellow fingers, Peer pressure, Chronic Disease, Fatigue, Allergy, Alcohol Consuming, all these 9 factors have a strong correlation with lung cancer. The use of these clinical factors to determine whether there is a risk of lung cancer has appeared less frequently in previous studies, providing partial support for people to prevent lung cancer by changing their lifestyle habits, and pointing out the direction for future in-depth research in this area.

Keywords: Lung cancer, pathogenic factors, binary logistic regression model.

1. Introduction

In recent years, more and more people worldwide have died prematurely due to cancer, and lung cancer is one of the leading causes of cancer-related deaths. In 2022, there were 19.96 million new cases of cancer worldwide, of which lung cancer accounted for 12.4% (2.48 million), making it the cancer with the highest number of new cases that year [1]. Lung cancer has a high fatality rate according to worldwide investigation, it was also the cancer with the highest number of deaths in 2022 (1.82 million, 18.7%) [1]. According to data from the Chinese Center for Disease Control and Prevention Weekly, lung cancer incidence and mortality rates in China are rising. In 2020, lung, colorectal, gastric, breast, and liver cancers accounted for 18%, 12%, 11%, 9%, and 9% of the nation's overall cancer cases, respectively, accounting for 59% of the total number of cancer cases in China [2].

Cancer has the characteristic of low difficulty in early cure, and detection and treatment in the early stages of cancer development are very important [3]. Research on cancer prediction methods has a long history. Methods for predicting lung cancer include screening for lung cancer susceptibility genes and constructing regression models, predicting based on metabolic profiles, and predicting based on miRNAs [4-6]. These methods can achieve accurate prediction of early lung cancer, but all require obtaining tissue samples from patients for analysis. The establishment of a cancer prediction method that can predict cancer risk based on clinical factors has the following three meanings: firstly, it has social significance. For society, predicting the risk of lung cancer through clinical factors can to some extent determine the high-risk population for lung cancer. Then, by conducting more accurate lung cancer related examinations on high-risk populations, lung cancer can be detected early and the mortality rate of lung cancer can be reduced; Then there is the significance for individuals: This low-cost, convenient prediction method helps individuals intuitively assess their lung cancer risk and identify key clinical factors, encouraging lifestyle changes to reduce that risk. Finally, regarding the significance of subsequent research, this study provides a low-cost and convenient method for predicting lung cancer, which provides a research direction for future studies.

According to the "Guidelines for Screening and Early Diagnosis and Treatment of Lung Cancer in China" and some literature, the main risk factors for lung cancer include smoking, passive smoking, alcohol consumption, etc., which are associated with an increased risk of lung cancer [7-9]. The probability of developing lung cancer is related to factors such as age, diet, and living and working environment. This study performed regression analysis on 23 factors, such as age, gender, environmental air pollution, and alcohol consumption, using 309 samples from Kaggle to develop a model for assessing lung cancer risk based on these factors.

2. Methods

2.1. Data Source

The lung cancer prediction data in this article is sourced from the Cancer Online Prediction website, and the raw data is saved in CSV format by MS NANCY AL ASWAD will be uploaded to the Kaggle website in 2022. This dataset provides important insights into the correlation between risk factors and the probability of developing lung cancer. The dataset includes 309 samples and 15 risk factors corresponding to each sample, as well as information on whether they have lung cancer [10].

2.2. Variable Introduction

The data used in this article contains 309 instances and 15 variables. Among them, there are 147 females and 162 males; 39 people were not diagnosed with cancer, 270 people were diagnosed with cancer. All 15 variables are represented in the Table 1.

Table 1. The relevant information of 15 variables

Term	Variable Name	Type	Range
Gender	X1	Categorical	0-male, 1-female
Age	X2	Numeric	21 to 87 years old
Smoking	X3	Categorical	0-NO, 1- YES
Yellow fingers	X4	Categorical	0-NO, 1- YES
Anxiety	X5	Categorical	0-NO, 1- YES
Peer pressure	X6	Categorical	0-NO, 1- YES
Chronic disease	X7	Categorical	0-NO, 1- YES
Chest pain	X8	Categorical	0-NO, 1- YES
Fatigue	X9	Categorical	0-NO, 1- YES
Allergy	X10	Categorical	0-NO, 1- YES
Wheezing	X11	Categorical	0-NO, 1- YES
Alcohol consuming	X12	Categorical	0-NO, 1- YES
Coughing	X13	Categorical	0-NO, 1- YES
Shortness of breath	X14	Categorical	0-NO, 1- YES
Swallowing difficulty	X15	Categorical	0-NO, 1- YES
Lung cancer	Y	Categorical	0-NO, 1- YES

2.3. Model Principle

The method used in this study was binary logit regression analysis of the dataset using SPSS. 15 factors mentioned in Table 1 as independent variables, with diagnosis of lung cancer as the dependent variable. Use logistic regression models to fit each variable for predictive diagnosis. The mathematical expression of logistic regression model is:

$$P\ (Y) = \frac{exp(\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_mX_m)}{1+exp(\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_mX_m)} \tag{1}$$

In the mathematical expression above, β_0 is a constant term, $\beta_1, \beta_2, \ldots, \beta_n$ is the regression coefficient corresponding to each independent variable. Using logic functions (i.e., sigmoid functions):

$$P\ =\ \partial(Z) = \frac{1}{1+e^{-Z}} \tag{2}$$

Where $Z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_m X_m$, then $L(p) = ln(\frac{p}{1-p})$. So, the logistic regression model can be expressed in the following linear form:

$$ln(\frac{p}{1-p}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_m X_m \tag{3}$$

3. Results and Discussion

3.1. Descriptive Statistic

Figure 1 shows the number of people corresponding to each category in each categorical variable in the dataset. As X2 is a numerical variable, it is not presented in the figure 1:

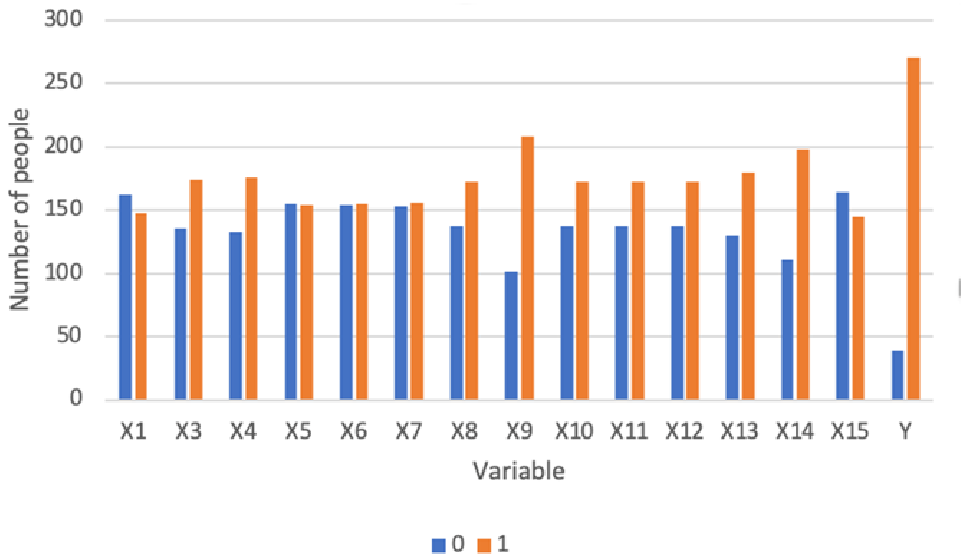

Fig. 1 The number of people corresponding to each category in the categorical variable

According to the information in Figure 1, there are a total of 15 independent variables, with a sample size of 309, which is approximately 20 times (more than 5 times) that of the independent variables; The positive cases in the outcome variable (Y) account for 87% (over 15%) of the total sample size, indicating that regression analysis can make good predictions for this dataset.

Figure 2 illustrates the number of individuals with and without lung cancer in the presence of various factors:

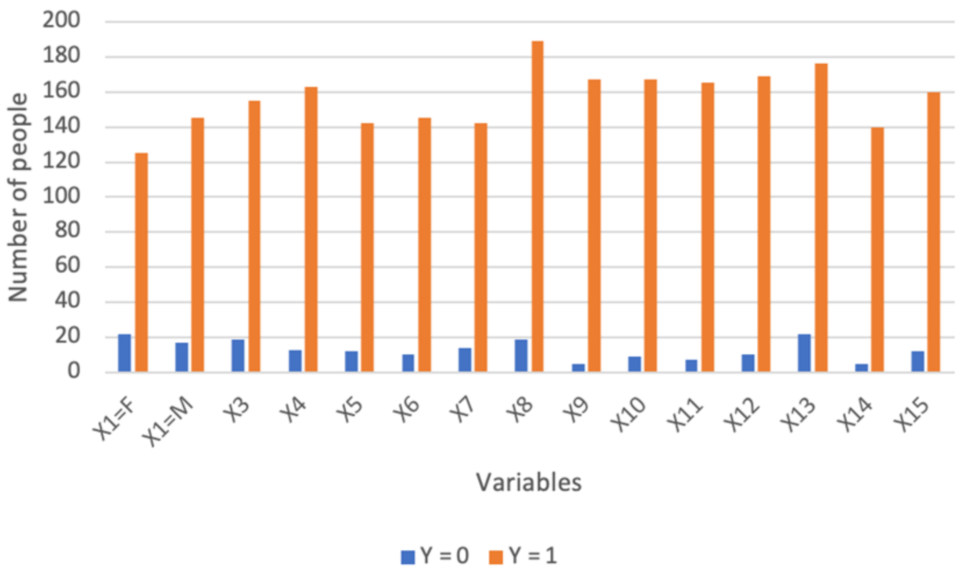

Fig. 2 The counts of individuals with and without lung cancer for each variable

From Figure 2, it can be seen that the ratio of cancer incidence in males (X1=M) to females (X1=F) in the sample is approximately 50:43, close to 1:1. Therefore, it can be inferred that gender is not related to the probability of lung cancer. Since X2 is a continuous numerical variable, it is divided into four age groups for statistical analysis. The number of people with and without lung cancer in each age group is shown in Figure 3:

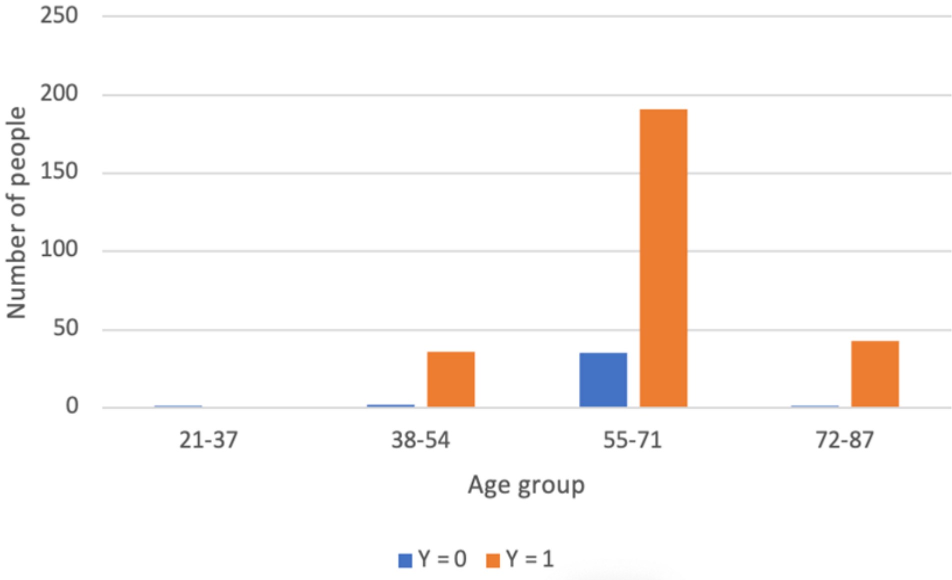

Fig. 3 Number of lung cancer patients corresponding to different age groups

From Figure 3, it can be seen that the number of cancer patients in the dataset generally increases with age, with the highest number of lung cancer patients in the 55-71 age group.

3.2. Logistic Regression Results

The factors for conducting logistic regression analysis include Gender, Age, Smoking, Yellow Fingers, Anxiety, Peer Pressure, Chronic Disease, Chest pain, Fatigue, Allergy, Wheezing, Alcohol Consuming, Coughing, Shortness of breath, Swallowing difficulty, all data in the dataset are used as training data, Adopting the LBFGS optimization algorithm. Table 2 shows the results of training the model:

Table 2. The prediction results of the model

Observed		Predicted		
		LUNG_CANCER		Percentage Correct
		NO	YES	
LUNG_CANCER	NO	29	10	74.4
	YES	7	263	97.4
Overall Percentage				94.5

According to Table 2, when predicting 39 samples that did not actually suffer from lung cancer, 29 samples were predicted to not have lung cancer, with an accuracy rate of 74.4%. When predicting 270 samples with actual lung cancer, 263 were predicted to have lung cancer, with an accuracy rate of 97.4% and a comprehensive accuracy rate of 94.5%. indicating that this model has high accuracy. Table 3 shows the results of binary logistic regression:

Table 3. The information of the variables in the equation

	B	S.E.	Wald	df	Sig	Exp(B)
X1	-0.526	0.709	0.551	1	0.458	0.591
X2	0.022	0.034	0.413	1	0.520	1.022
X3	1.776	0.702	6.402	1	0.011	5.906
X4	1.376	0.743	3.436	1	0.064	3.961
X5	0.888	0.813	1.193	1	0.275	2.430
X6	1.731	0.660	6.875	1	0.009	5.648
X7	3.192	0.888	12.909	1	0.000	24.326
X8	3.070	0.825	13.844	1	0.000	21.551
X9	1.646	0.769	4.583	1	0.032	5.187
X10	0.966	0.834	1.342	1	0.247	2.628
X11	1.410	0.799	3.114	1	0.078	4.095
X12	3.311	1.072	9.547	1	0.002	27.420
X13	-0.729	0.760	0.920	1	0.338	.482
X14	3.122	1.130	7.636	1	0.006	22.694
X15	0.559	0.689	0.658	1	0.417	1.749
Constant	-8.338	2.507	11.061	1	0.001	.000

From Table 3, it can be seen that the significance (sig) of X1, X2, X5, X10, X13 and X15 is much higher than 0.05, indicating that these 6 variables have little relationship with lung cancer. The B column in Table 3 represents the β value corresponding to each independent variable, and constant represents β_0. Therefore, the following mathematical expression can be obtained:

$$ln\left(\frac{p}{1-p}\right) = -8.338 - 0.526X_1 + 0.022X_2 + 1.776X_3 + \cdots + 0.559X_{15} \quad (4)$$

4. Conclusion

Current research shows that among 15 elements, only Smoking, Yellow Fingers, Peer Pressure, Chronic Disease, Fatigue, Allergy, Alcohol Consumption, Coughing, and Swallowing Difficulty have a strong association with lung cancer. These nine elements effectively predict lung cancer risk, achieving a high accuracy of 94.5% in both the test and training sets. Analyzing the Exp (B) value of the data can reveal factors that have a significant impact on individual lung cancer. This can provide better advice for people to prevent lung cancer, while also making them aware of their potential risk of lung cancer and adjusting their lifestyle habits in a timely manner. This study only used a dataset containing 309 samples, and most of the variables in the dataset were categorical variables. In order to further investigate, more relevant datasets with continuous variables should be collected. In the future, this study needs to collect more data and compare results, while also using more advanced models to analyze potential interactions and nonlinear effects between variables.

References

[1] Global cancer statistics. GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries - Bray - CA: A Cancer. Journal for Clinicians - Wiley Online Library, 2022.

[2] Kadir T, Gleeson F. Lung cancer prediction using machine learning and advanced imaging techniques. Translational lung cancerresearch, 2018, 7(3): 304.

[3] Loud J T, Murphy J. Cancer Screening and Early Detection in the 21st Century. Semin Oncol Nurs, 2017, 33(2): 121-128.

[4] Hu Huating. Screening of lung cancer susceptibility genes and construction of lung cancer prediction model. Anhui University of Finance and Economics, 2019.

[5] Li Chenwei. Construction of Nomogram Model for Early Lung Cancer Prediction Based on Metabolic Spectrum. Liaoning: Dalian Medical University, 2022.

[6] Nan Juan. The enormous potential of miRNAs as cancer prediction and prognostic markers. Chinese Journal of Lung Cancer, 2013.

[7] Hejie, Li Ni, Chen Wanqing, et al. Guidelines for Screening and Early Diagnosis and Treatment of Lung Cancer in China (2021, Beijing). Chinese Oncology, 2021.

[8] Wang K, et al. The culprit of cancer causing death-smoking. Chinese Journal of Pulmonary Disease (Electronic Edition), 2015, 8(3): 63.

[9] Liu Yujie. Assessment of Death Risk and Health Economic Losses of Cancer Patients Related to Air Pollution in Chengdu City. Chengdu Medical College, 2022.

[10] Hong Z Q, Yang J Y. Optimal Discriminant Plane for a Small Number of Samples and Design Method of Classifier on the Plane. Pattern Recognition, 1991, 24: 317-324.