The Research of Influence Factors that Possibly lead to Lung Cancer

Yun Dong[1, *, †], Zhitong Guo[2, †] and Peixian Hu[3, †]

[1] School of Mathematical Sciences, Nankai University, Tianjin, 300110, China

[2] School of Mathematics, Hunan University, Changsha, 410006, China

[3] Ningbo Binhai international Cooperative School, Ningbo, 315800, China

* Corresponding Author Email: 2011173@mail.nankai.edu.cn

†These authors contributed equally.

Abstract. Although previous studies have demonstrated that the incidence of lung cancer is associated with genetic predisposition, smoking, and air quality, there are still many unidentified factors that deserve to be studied. In this research, the method Binary Probit Model is used to deal with the data from the online Lung Cancer Prediction System which was published and updated in 2022 for 309 individuals. It is concluded that although lung cancer has no connection with Gender, Age, Anxiety, Wheezing, Alcohol Consuming, Shortness of Breath, or Chest Pain, it has a relatively strong connection with Yellow Finger and Allergy and a particularly strong relationship with Smoking, Peer Pressure, Swallowing Difficulty, Chronic Disease, Fatigue, and Coughing. Many of the factors having a possibly strong connection with cancer have never appeared in previous studies, such as Peer Pressure, Swallowing Difficulty, Chronic Disease, and Fatigue, which provides some new perspectives to study the pathogenesis and treatment of lung cancer and point to the way for further research afterward.

Keywords: Lung cancer; pathogenic factors; binary probit model.

1. Introduction

Lung cancer had become one of the most valued diseases. Bray et al. analyzed the data on Age-SIR(S) of the World Population of lung cancer in 2018 and concluded that lung cancer had the highest incidence and mortality rates in the world [1]. Deng also concluded that lung cancer mortality had been the top global MT and it had become leading cancer globally [2]. In recent decades, the incidence and mortality of lung cancer in China had been on the rise [3, 4]. Therefore, it is important to understand the factors contributing to lung cancer to detect and treat the disease as early as possible and reduce mortality. This paper aims to study different potential factors that possibly lead to lung cancer, and help people to assess their risk of lung cancer and take meaningful preventive measures according to the risk level.

The factors involved in the development of lung cancer were very complex, and its occurrence was the result of multi-factorial, multi-gene, and multi-step interactions [5]. Domestic and foreign scholars had found some correlation between lung cancer and smoking [6], passive smoking [7], and sleep duration [8, 9]. In addition, Sun found that other factors also had some impact on lung cancer [10]. However, this literature studied fewer lung cancer factors and lacked systematic and complete statistics. Therefore, this paper focuses on 15 factors (Gender, Age, Smoking, Yellow fingers, Anxiety, Peer-pressure, Chronic Disease, Fatigue, Allergy, Wheezing, Alcohol, Coughing, Shortness of Breath, Swallowing Difficulty, and Chest pain) to study whether they have any effect on lung cancer, and further select an appropriate model to study the degree of correlation between these factors and lung cancer.

In similar directions, Wang et al. used a model of Population Attribution Fraction (PAF) [11], but the data in the literature had limitations, in which the parts of the data that were difficult to obtain were replaced with reasonable assumptions. The accuracy of some of the data needed to be improved, at the same time, the literature excluded public places, which was an important source of secondhand smoke. Tian et al. used a model of Logistic Regression Analysis [12], which could efficiently and concisely output well-calibrated predicted probabilities. It required a case-control study in medical research to establish multiple matched groups, usually 1 case and 1 or 2 or 3 control matches. However, this model had high requirements for data and was not as accurate when the sample size of the control group (without lung cancer) was small. Liu et al. constructed a model of Unconditional Logistic Regression [13] but omitted some easily ignored factors other than air pollution and smoking, and the model had fewer application cases that could be used as additional learning and research objects. Feng used a model with Matching Matrix and Gray Correlation Analysis [14], but only considered the spatial and temporal relationship between environmental factors and lung cancer, without examining more factors contributing to lung cancer incidence. Qiao et al. used a model with Multi-factor Poisson Regression Analysis [15], but the majority of the screening population was male, and the possibility of the non-smoking population participating in screening was not fully considered. In addition, the model was not suitable when the data were too discrete.

In summary, after consideration and optimization, this paper will use the Probit Regression model to study the effect of these 15 factors on lung cancer, i.e., whether they are factors that possibly lead to lung cancer.

2. Methods

2.1. Data Sources

The data for this literature is collected from the Kaggle website, which was compiled by Mysar Ahmad Bhat from the online Lung Cancer Prediction System and published and updated in 2022 for 309 individuals.

2.2. Variable Selection

The data used in this paper count a total of 309 people, including those who have and do not have lung cancer, of whom 162 are male and 147 are women. The patients' ages range from 21 to 87 years. The data contains 15 variables (Gender, Age, Smoking, Yellow fingers, Anxiety, Peer-pressure, Chronic Disease, Fatigue, Allergy, Wheezing, Alcohol, Coughing, Shortness of Breath, Swallowing Difficulty, and Chest pain).

Table 1. Logogram and numbers of the 15 factors

Elements	Logogram	Number1	Cancer1
Gender	x_1	309	270
Age	x_2	309	270
Smoking	x_3	174	155
Yellow Finger	x_4	176	163
Anxiety	x_5	154	142
Peer Pressure	x_6	155	145
Chronic Disease	x_7	156	142
Fatigue	x_8	208	189
Allergy	x_9	182	167
Wheezing	x_{10}	172	163
Alcohol Consuming	x_{11}	172	165
Coughing	x_{12}	179	169
Shortness of Breath	x_{13}	198	176
Swallowing Difficulty	x_{14}	145	140
Chest Pain	x_{15}	172	160

*Number 1: The number of people suffering from the disease.
**Cancer1: The number of lung cancer patients suffering from the disease.

Table 1 shows the number of people and lung cancer patients who have the disease. As shown in Table 1, for the convenience of writing, the logogram of the factors is as above. The sample of data is 309, of which 270 have lung cancer.

Table 2. Gender distribution by age group

Age	[21-30]	[31-40]	[41-50]	[51-60]	[61-70]	[71-80]	[81-90]
Number2	1	2	12	108	132	51	3
Female	1	1	6	53	56	27	3
Male	0	1	6	55	76	24	0
Cancer2	0	2	10	92	114	50	2

*Number 2: The number of people in each age group.

**Cancer 2: The number of people in each age group who have lung cancer.

Table 2 shows the number of females and males, and lung cancer patients in each age group. The majority of people in the data were aged 51-60 and 61-70 years old.

2.3. Research Protocol

This paper uses the Probit Regression model, whether or not to have lung cancer is the dependent variable(Y), and the 15 factors are the independent variables(X), where 0 represents no and 1 represents yes. Next, this paper uses SPSSAU and SPSSPRO to analyze the relationship between the effect of X on Y, i.e., the relationship between the 15 factors on lung cancer.

2.4. Model Principle

Probit regression is used for statistical analysis of data whose dependent variable is a categorical variable, which is similar to Logistic regression. Deforming Logistic mode, it can be seen that:

$$P = \frac{e^{X\beta}}{1+e^{X\beta}} \tag{1}$$

The right-hand side of the above equation ($\frac{e^{X\beta}}{1+e^{X\beta}}$) happens to resemble the probability distribution function of the standard growth distribution (also known as the Logistic distribution). The Probit model assumes that the probability distribution function on the right is similar to that of the standard normal distribution:

$$P = \int_{-\infty}^{X} \frac{1}{\sqrt{2\pi}} e^{-\frac{(X\beta)^2}{2}} \tag{2}$$

2.5. Model Testing

It can be seen from Table 3 that as model fitting quality is judged by model prediction accuracy, the overall prediction accuracy of the research model is 94.82%, and the model fitting is acceptable. When the true value is 0, the prediction accuracy is 74.36%. In addition, when the true value is 1, the prediction accuracy rate is 97.78%.

Table 3. Binary probit regression prediction accuracy summary

		predicted value		forecast accuracy	Prediction error rate
		0	1		
true value	0	29	10	74.36%	25.64%
	1	6	264	97.78%	2.22%
Summary				94.82%	5.18%

3. Results and Discussion

Figure 1 shows a series of factors that cause lung cancer in the proportion of people with lung cancer. These factors that cause lung cancer are Smoking, Yellow fingers, Anxiety, Peer pressure, Chronic disease, Fatigue, Allergy, Wheezing, Alcohol, Coughing, Shortness of Breath, Swallowing difficulty, and Chest pain. There are a total of 309 samples, the proportion of lung cancer samples that were identified from the samples and classified according to different influencing factors was determined to determine the probability that each risk would cause lung cancer.

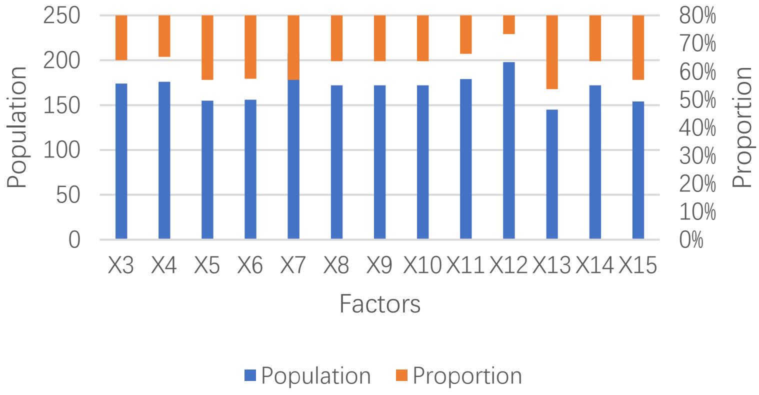

Fig. 1 Factors in the proportion of people with lung cancer

In above Figure 1, it is not difficult to see that fatigue accounts for the highest proportion, up to 77%. Shortness of Breath had the lowest share at 53.7 percent. Thus, according to the chart data, fatigue accounts for the highest proportion, leading to the greatest hidden risk of lung cancer.

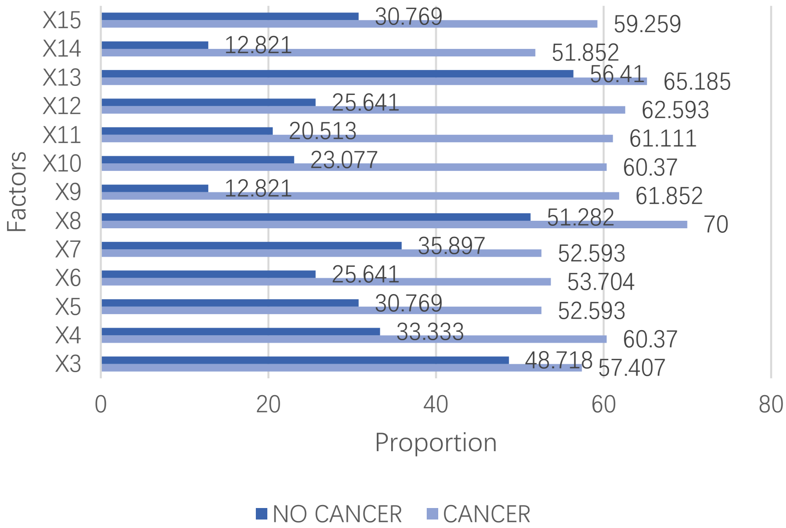

Fig. 2 Comparison of cancer and non-cancer under different factors

Based on the comparison chart, the difference between the proportion of getting cancer and the proportion of no cancer on the factors x_3- x_{15} is relatively large with the least disparity accounting for more than 10%, leading to the greatest hidden risk of lung cancer, which is worth studying and analyzing.

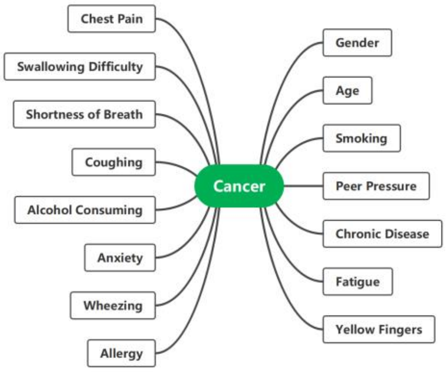

Fig. 3 Variable-related schematic

As shown in Figure 3, the research introduced various factors that might be related to lung cancer into the model, including Gender, Age, Smoking, Yellow Fingers, Anxiety, Peer Pressure, Chronic Disease, Fatigue, Allergy, Wheezing, Alcohol Consuming, Coughing, Shortness of Breath, Swallowing Difficulty, and Chest Pain. Through calculations, this study yields the final linear regression equation:

$$Probit(p) = -4.321 - 0.266x_1 + 0.010x_2 + \ldots + 0.291x_{15} \tag{3}$$

Where p represents the probability of Lung Cancer being 1.

Table 4. Model results

Elements	regression coefficient	standard error	z value	p-value	95% CI	marginal effect
x_1	-0.266	0.39	-0.681	0.496	-1.030 ~ 0.499	-0.021
x_2	0.010	0.018	0.574	0.566	-0.025 ~ 0.046	0.001
x_3	1.022	0.392	2.605	0.009	0.253 ~ 1.791	0.091
x_4	0.808	0.406	1.991	0.046	0.013 ~ 1.604	0.065
x_5	0.486	0.444	1.095	0.273	-0.384 ~ 1.356	0.038
x_6	1.003	0.364	2.753	0.006	0.289 ~ 1.717	0.079
x_7	1.791	0.498	3.599	0.000	0.816 ~ 2.766	0.131
x_8	1.756	0.447	3.926	0.000	0.879 ~ 2.632	0.164
x_9	0.900	0.421	2.138	0.033	0.075 ~ 1.726	0.081
x_{10}	0.557	0.454	1.228	0.219	-0.332 ~ 1.446	0.048
x_{11}	0.817	0.441	1.854	0.064	-0.047 ~ 1.681	0.076
x_{12}	1.819	0.581	3.131	0.002	0.680 ~ 2.957	0.178
x_{13}	-0.425	0.413	-1.030	0.303	-1.234 ~ 0.384	-0.033
x_{14}	1.720	0.626	2.748	0.006	0.493 ~ 2.947	0.128
x_{15}	0.291	0.369	0.788	0.431	0.433~1.014	0.024
Constant term	-4.321	-	-	-	-	-

*Dependent variable: Lung Cancer

It is manifest from Table 4 that the model gets the conclusion that by judging whether the p-value is greater than 0.05, this study can further infer what factors may be associated with lung cancer. On the one hand, the p values of Gender, Age, Anxiety, Wheezing, Alcohol Consuming, Shortness of Breath, and Chest Pain are all greater than 0.05, meaning they do not affect lung cancer. On the other hand, the p values of Yellow Finger and Allergy are all smaller than 0.05 but larger than 0.01, which shows that these factors have relatively strong affection on lung cancer and the p values of Smoking, Peer Pressure, Swallowing Difficulty, Chronic Disease, Fatigue, and Coughing are less than 0.01, which illustrate that these factors have a particularly strong relationship with lung cancer. When it comes to the influence relations, regression coefficient value combined with Marginal effect value show the specific influence factors: the change (increasing) in lung cancer will be 254.79%, 203.34%, 309.01%, 443.28%, 410.16%, 202.51%, 314.06%, 337.73% with one unit increase of Smoking, Yellow Fingers, Peer Pressure, Chronic Disease, Fatigue, Allergy, Coughing, Swallowing Difficulty separately. There is no denying that these factors have a significant impact on the likelihood of getting lung cancer, as these data are all large comparably, which means they have some reference and research value.

Compared to previous studies, they tend to use specific analysis of a single factor, which is limited to known possible factors for developing lung cancer, such as genetic inheritance, smoking (including secondhand smoke), and air quality of the patient's living environment. In contrast, the factors covered in this study are more extensive and comprehensive, which can not only effectively avoid the errors caused by not controlling for a single variable, but also broaden the ideas of future research on lung cancer, helping medical staff to identify more directions for treatment and detect lung cancer earlier so that timely treatment can be obtained.

4. Conclusion

The current study selects diverse data and focuses on influencing factors that may be associated with developing lung cancer. It was concluded that having lung cancer may be related to Smoking, Yellow Fingers, Peer Pressure, Chronic Disease, Fatigue, Allergy, Coughing, and Swallowing difficulties, most of which have not been paid attention to before.

It cannot be denied that due to the limited amount of data, this model may have errors in addition to the factors, and the sample did not cover all ages and ethnicities, causing possible differences, which may also affect the accuracy of the results. However, the study still has a lot of value and merits. First of all, the approach taken in this study is innovative. On the one hand, a graphical approach was used to visually analyze the differences in the proportion of each factor in the population with lung cancer and the population without lung cancer. This enables visualization of the experiment and makes the results clearer and more intuitive. On the other hand, instead of using a single-factor analysis method as in many previous experiments, multiple linear regression was chosen for this experiment, making it more comprehensive. Secondly, it has some positive effects on lung cancer treatment. In addition to the factors known to be associated with cancer such as Smoking, Chronic Disease, more factors may be related to lung cancer that deserve attention and consideration, such as Coughing, Swallowing Difficulty, Peer Pressure, and so on. Whether these factors are associated with the development of lung cancer requires further medical investigation, which means that these findings will point the way to further relevant research in the future. Once a new causative factor other than the one already identified is discovered, this can help us detect lung cancer earlier, and treat it as soon as possible, improving survival rates and patients' quality of life.

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: a cancer journal for clinicians, 2018, 68(6).

[2] Deng Dajun. World Cancer Report 2020 edition is released--Adapting cancer prevention measures to new trends in cancer prevalence. Journal of Multidisciplinary Cancer Management (Electronic Version), 2020, 6(03): 27-32.

[3] Chen Wanqing, Sun Kexin, Zheng Rongshou, et al. Cancer incidence and mortality in China,2014. Chinese Journal of Cancer Research, 2018, 30(01): 1-12.

[4] Li Xiang, Gao Shen. Analysis of lung cancer incidence, prevalence, and mortality trends among Chinese residents from 1990-2019. Chinese Journal of Prevention and Control of Chronic Diseases, 2021, 29(11): 821-826.

[5] Lu Wanting, Chen Minghui, Huang Liping, et al. Relationship between behavior and lifestyle and the development of lung cancer. Strait Journal of Preventive Medicine, 2018, 24(04): 19-22.

[6] Li Jinghan, Qi Yingjia. Research progress in the development factors, diagnostic modalities and treatments of lung cancer. World Latest Medicine Information, 2019, 19(28): 123-125.

[7] Öberg M, Jaakkola M, Woodward A, et al. Worldwide burden of disease from exposure to second-hand smoke: a retrospective analysis of data from 192 countries. Lancet, 2011, 377(9760): 139-146.

[8] Bai Yansen, Li Xiaoliang, Wang Suhan, et al. A cohort study of retired workers' nighttime sleep duration and risk of lung cancer. Proceedings of the National Conference on Tumor Epidemiology and Tumor Etiology, 2015, 156-157.

[9] Schernhammer E, Feskanich D, Liang G, et al. Rotating night-shift work and Lung cancer risk among female nurses in the United States. American Journal Epidemiology, 2013, 178(9): 1434-1441.

[10] Sun Weixiu. Meta and Bioinformatics analysis of the relationship between smoking and lung cancer development. Supervised by Ren Z.L. Dalian Medical University, 2020.

[11] Wang Jianbing, Jiang Yong, Wei Wenqiang, et al. Estimation of cancer incidence and mortality attributable to smoking in China. Cancer Causes & Control, 2010, 21(6): 959-965.

[12] Tian Chenhong, Zhu Liping, Liao Qiaoying. Investigation of major risk factors for lung cancer development in rural areas. Chinese Journal of Public Health Management, 2015, 31(06): 894-895.

[13] Liu Liqun, Liu Xiaoyan, Ma Xiangyun, et al. Analysis of the associations of indoor air pollution and tobacco use with morbidity of lung cancer in Xuanwei, China. Science of The Total Environment, 2019, 135232.

[14] Feng Ziyu. Spatial and temporal distribution of PM_(2.5) and lung cancer incidence in China and its association characteristics analysis. East China Normal University, 2021.

[15] Qiao Liang, Liu Mingyan, Zhou Peng, et al. Screening results and factors influencing the incidence of lung cancer in a high-risk cohort in Longquanyi District, Chengdu City, 2014-2018. China Cancer, 2022, 31(12): 975-982.