Dean&Francis

ISSN 2959-6157

Multifunctional Analysis of Lung Cancer Risk-Interactions Between Research Genetic, Environmental, and Lifestyle Factors

Xinyan Wang

Shijiazhuang No.5 Middle School, Shijiazhuang, China
ayhhxinyan@gmail.com

Abstract:

As the core organ of the respiratory system, the lungs maintain acid-base balance through gas exchange, play a defensive role for the organism, and are involved in metabolism and secretion, making them an important organ for sustaining life activities; however, nowadays, there are more and more patients with lung cancer, which is a serious threat to human health. In this study, the researchers investigated the effects of seven factors, namely age, gender, air pollution, smoking, obesity, alcohol use, and dust allergy, on lung cancer by analysing biomedical, environmental, personal behavioural, and lifestyle factors based on the Kaggle dataset (which contains 999 clinical cases) using logistic regression and stepwise regression analyses. The results showed that dust allergy had a significant positive effect on lung cancer risk, while air pollution, obesity, and smoking. The researcher negatively associated, contradicting traditional epidemiological findings. The model explained 53.5% of the variance, emphasising the synergistic effect of environmental and immune abnormalities, warning, and providing a new direction for prevention strategies.

Keywords: Linear regression analysis; Stepwise regression analysis; Histogram analysis

1. Introduction

Cancer is one of the most important diseases endangering human health and life, and lung cancer, as the malignant tumour with the highest incidence and mortality rate in the world, poses a serious threat to human survival. In recent years, the incidence of lung cancer has continued to rise, and although the global mortality rate of lung cancer has declined with the deepening of etiological research and the advancement of precision medicine technology, the danger of lung cancer is still serious [1].

According to data from the China Tumour Registry, China ranked first in lung cancer incidence and

ISSN 2959-6157

mortality rates in 2016. From 1990 to 2021, the age-standardised incidence rate of lung cancer in China showed an overall increasing trend [2]. At this stage, the pathogenesis of lung cancer has not been fully elucidated, and it is widely believed that its development is closely related to tobacco exposure, genetic susceptibility, environmental pollution, and other factors. In China, the prevention and control of lung cancer is a key task of the medical and healthcare system, and due to the differences in epidemiological characteristics of different regions, prevention and control strategies need to be formulated in accordance with the actual local situation [3].

Age-standardized incidence rates (ASIRs), mortality rates, and disability-adjusted life-years (DALYs) rates for lung cancer among adolescents and young adults (AYAs) in China generally show a decreasing trend in 1990-2021, but at a rate lower than the global level. The rates of ASIRs, ASMRs, and DALYs for lung cancer in Chinese AYAs turn up in 2013 to 2021. Ageing is the main cause of the rising lung cancer burden in China, while globally it is mainly driven by population growth [4]. Smoking, PM2.5, secondhand smoke, and indoor pollution are the main risk factors. Despite the positive global trend, non-smokers have a relatively increased burden of lung cancer and need to be targeted for prevention and control. Climate change significantly enhances the carcinogenic effects of atmospheric pollutants. In addition, the aerosol released by e-cigarettes contains the mutagenic substance formaldehyde glycerol, which has been shown to significantly increase the Ki-67 proliferation index of alveolar type II epithelial cells [5]. Therefore, the key prerequisite for researching lung cancer treatment methods lies in identifying its influencing factors; in-depth investigation of these factors is of great significance for advancing treatment strategies.

This study analyses the impact of factors such as patient age, gender, living environment, and lifestyle on the risk of developing lung cancer, thereby providing a basis for further exploration of treatment methods for lung cancer.

2. Method

2.1 Data Sources and Description

The source of this data is Kaggle [6]. This dataset comes from cancer patient data sets derived from The Devastator study, which summarizes and classifies 462,000 patients in China by collecting them in groups with an average of more than six years of lung cancer. With 999 clinical examples, this table has a wide range of research scope, which is universal and practical, and the completeness, credibility, and compatibility of the table are all 100%, which has a certain degree of credibility.

In the original dataset, the researcher included 7 variables, including gender, age, smoking, air pollution, obesity, alcohol use, and dust allergy. For data preprocessing, this paper coded dichotomous and multicategorical variables as factors, and the corresponding representation is shown.

2.2 Variable Selection

In this study, the researchers used decomposition analysis, which is a term used to quantify the independent contribution of different factors to overall change or variation, to control the variables of different factors to explore the incidence and severity of lung cancer, with the total change being the sum of several components. Also, constructing a model using logistic regression binary logistic regression

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \alpha \quad (2)$$

In this formula, y represents the severity of lung cancer, β denotes the unstandardised coefficient corresponding to the factor, and x represents various relevant factors.) to explore the factors influencing lung cancer. In addition, this paper used visual analysis to carefully examine the impact of various factors at different levels on lung cancer. Moreover, the researcher conducted to investigation into the mechanism of carcinogenesis of different substances and to prevent cancer from an environmental aspect.

3. Result and Discussion

3.1 Linear Regression and Stepwise Regression

Table 1 uses regression analyses to examine the relationship between factors influencing lung cancer and the severity of lung cancer, whether there is an influential relationship, and the degree of influence.

Table 1. Linear regression analysis results

	Non-standardised coefficient		Standardised coefficient	t	p	Collinearity diagnosis	
	B	standard error	$Beta$			VIF	Tolerance
constant	2.963	0.357	-	8.292	0.000**	-	-
Age	0.007	0.005	0.108	1.504	0.136	1.106	0.904
Gender	-0.127	0.129	-0.077	-0.989	0.325	1.276	0.784
Air Pollution	-0.191	0.046	-0.439	-4.130	0.000**	2.409	0.415
Alcohol use	-0.007	0.048	-0.022	-0.148	0.883	4.787	0.209
Dust Allergy	0.178	0.056	0.421	3.176	0.002**	3.750	0.267
Obesity	-0.158	0.039	-0.394	-4.024	0.000**	2.047	0.489
Smoking	-0.133	0.029	-0.391	-4.630	0.000**	1.522	0.657
R^2	0.569						
Adjustment R^2	0.536						
F	F (7,92)=17.328,p=0.000						
D-W value	1.817						
Note: Dependent variable Level							
* p<0.05 ** p<0.01							

Continue Table 1.

Table 2. Stepwise regression analysis

Stepwise regression analysis results (n=100)							
	Non-standardised coefficient		Standardised coefficient	t	p	Collinearity diagnosis	
	B	standard error	$Beta$			VIF	Tolerance
constant	2.956	0.166	-	17.860	0.000**	-	-
Air Pollution	-0.176	0.041	-0.406	-4.356	0.000**	1.847	0.541
Dust Allergy	0.180	0.043	0.425	4.139	0.000**	2.247	0.445
Obesity	-0.161	0.039	-0.402	-4.118	0.000**	2.030	0.493
Smoking	-0.132	0.027	-0.388	-4.863	0.000**	1.358	0.736
R^2	0.554						
Adjustment R^2	0.535						
F	F (4,95)=29.517, p=0.000						
D-W value	1.781						
Note: Dependent variable= Level							
* p<0.05 ** p<0.01							

Linear regression analysis with age, gender, air pollution, alcohol use, dust allergy, obesity, smoking as the independent variables and level as the dependent variable yields table 1 because the DW value is 1.817, which is approximately equal to 1.900, so the samples are independent, and the VIFs are all less than 5, so there is there is no multicollinearity, so the linear regression model can be used. From Table 2, this linear regression model is

$$Level = 2.963 + 0.007 * Age - 0.127 * Gender - 0.191 * AirPollution - 0.007 * Alcoholuse + 0.178 * DustAllergy - 0.158 * Obesity - 0.133 * Smoking \quad (1)$$

Since the R-squared value is 0.569, the above factors explain 56.9% of the influences because F (7,92)=17.328,p=0.000, which is less than 0.005, so at least one of the above factors will affect lung cancer. The

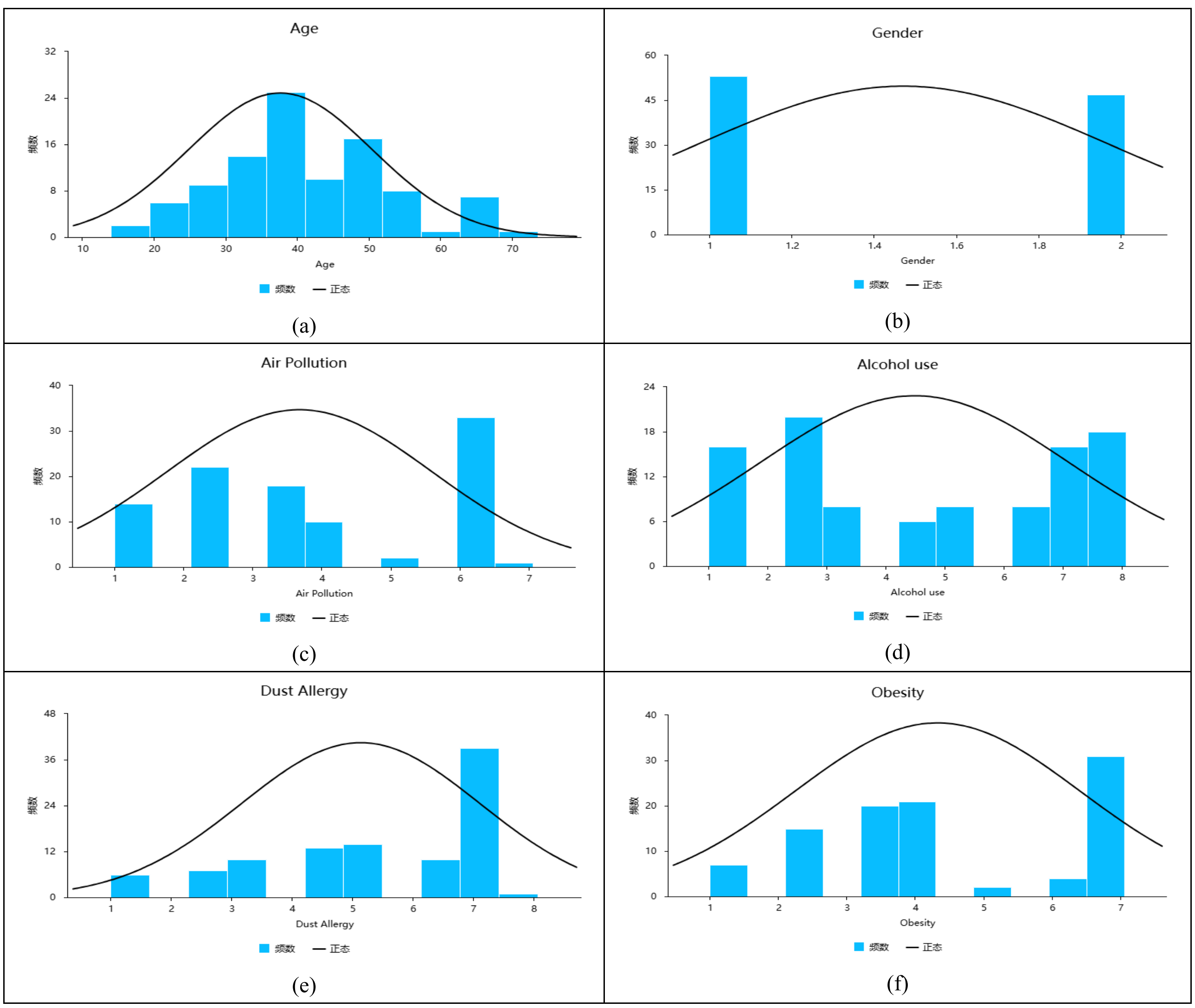
ISSN 2959-6157

significance of age, gender, and alcohol use is greater than 0.05, and the significance of the rest of the factors is less than 0.05, so age, gender, and alcohol have no significant effect on lung cancer, the regression coefficient value of air pollution, obesity and smoking is greater than 0, so air pollution, obesity and smoking have a significant negative effect on lung cancer, and the regression coefficient value of dust allergy is greater than 0, so dust allergy has a significant positive effect on lung cancer.

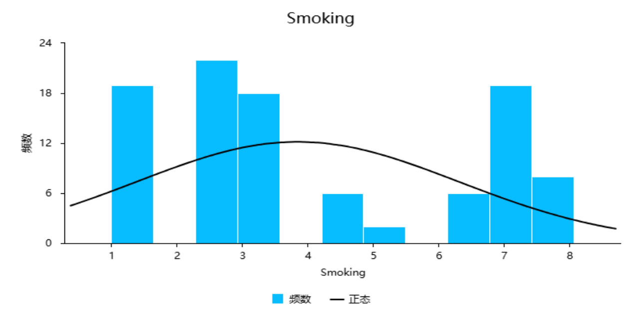
The omnivariate model had a certain analytical effect, but the researcher still had some non-significant variables. After stepwise regression, from table 2, the significant variables including air pollution, dust allergy, obesity, and smoking The researcher filtered out, and the overall model was significant with $F(4,95)=29.517$, $p=0.000$, and the adjusted R^2 was 0.535, which indicated that these four types of variables jointly explained 53.5% of the dependent variable, and the Durbin-Watson value (1.781) was close to 2, with strong linear regression model characteristics. The Durbin-Watson value (1.781) was close to 2, characterising a strong linear regression model.

Among them, dust allergy contributed the most, and its positive association may reflect the key role of the immune-inflammatory pathway in lung cancer imaging. Whereas air pollution, obesity and smoking The researcher all negatively associated, the negative coefficient of air pollution may imply the existence of some compensatory protective mechanism in highly polluted environments, the negative association of obesity may be related to the complex effects of metabolic disorders, and the negative coefficient of smoking contradicts the conventional epidemiological findings, which may reflect the sample specificity or the existence of a non-linear relationship.

3.2 Visual Analysis-Histogram



(a) (b) (c) (d) (e) (f)



(g)

Fig. 1 Histogram of All Variables, (a) represents age, (b) represents gender, (c) represents air pollution, (d) represents alcohol use, (e) represents dust allergy, (f) represents obesity, (g) represents smoking (Photo/Picture credit: Original).

Fig. 1 shows the histograms of seven variables. From the analysis of the figure, it can be seen that there is little difference in the binary variable of gender. Among the multi-category variables, the age range is from 0 to 70 years old, which is relatively consistent with a normal distribution; in the air pollution variable, there are seven levels in total. At level 5, the impact is very minor; at level 6, the impact increases sharply; and at higher levels, the impact is almost negligible. In the alcohol use variable, there are eight levels in total. The intermediate levels have a smaller impact, while levels 1, 2, 7, and 8 have approximately the same impact and are relatively significant; In the dust allergy variable, there are 8 levels, with mild allergies having a small impact on lung cancer, while severe allergies have a sudden increase in impact; in the obesity variable, there are 7 levels, with levels 5 and 6 having a small impact, while the remaining 5 levels have a noticeable impact, especially when severely obese, which has a significant impact on lung cancer; In the smoking variable, there are 8 levels. Except for levels 4, 5, and 6, which have a smaller impact than the average, both mild and severe smoking have a severe impact on lung cancer.

3.3 Discussion

Overall, the age-standardized mortality rate and daily rate for lung cancer attributable to tobacco smoking showed a downward trend globally and in the United States from 1990 to 2021, while the disease burden of lung cancer attributable to tobacco smoking in China showed an upward trend [7]. However, after quitting smoking, the risk of developing lung cancer decreases but never reaches the baseline level of non-smokers [8]. When analyzing the obesity paradox, the use of metformin is an important confounding factor. Compared to metabolically healthy adults, obese individuals generally exhibit insulin resistance, resulting in an increased risk of diabetes. Additionally, statins, aspirin, and other medications are contributing factors to this paradox, suggesting that the negative correlation between obesity and lung cancer may be related to the limited sample size of the study population [9]. Between 1990 and 2021, the PAF value for lung cancer attributable to air pollution in China significantly decreased, from 46.23% in 1990 to 32.66% in 2021, indicating a relative decrease in the contribution of air pollution to lung cancer. However, the trends in standardized mortality rates and standardized DALY rates, which both showed an initial increase followed by a decrease, reflect China's efforts in air pollution control and public health interventions, which have, to some extent, mitigated the rising rate of disease burden [10]. This study is constrained by its data sources, with the sample cohort exhibiting a high degree of homogeneity in demographic characteristics and clinical status, thereby imposing certain limitations on the findings. Future research should incorporate broader population samples and collect multidimensional data to comprehensively investigate the various factors associated with the occurrence and progression of lung cancer.

4. Conclusion

This study addresses the core question raised in the introduction, namely that the pathogenesis of lung cancer is the result of complex interactions between genetic, environmental, and lifestyle factors. Through empirical analysis, this study validates the importance of these factors while revealing findings that contradict traditional beliefs, thereby providing new directions for discussion in lung cancer etiology research. Among the four major factors influencing lung cancer, dust mite allergy has the strongest positive impact, while air pollution exhibits a negative impact, necessitating further verification of whether environmental adaptation mechanisms exist. Obesity and smoking are both negatively correlated with lung cancer, challenging traditional epidemiological conclusions, which may reflect

ISSN 2959-6157

sample specificity or nonlinear relationships. This study directly addresses the article's theme of lung cancer risk diversity. The findings emphasize the synergistic effects of environmental exposure and immune abnormalities, while also suggesting the complex roles of metabolic and behavioral factors, providing evidence for precision prevention strategies. For public health authorities, priority should be given to controlling air pollution and allergen exposure, especially in highly polluted areas. For clinical researchers, the "protective" association between obesity and smoking should be further explored to avoid misinterpretation. Policy makers should pay attention to the increasing risk of lung cancer. They should focus on the rising trend of lung cancer risk among non-smokers and adjust the priorities of prevention and control. Lung cancer is not only a medical issue but also a response to environmental and social governance. Exploring its influencing factors can not only save lives but also serve as a warning for us to prioritize health.

References

[1] Li S., Liu Y., Zhou J., Wang D., Sun Z., Chen R. Analysis of physical activity levels and related factors in patients undergoing initial chemotherapy for lung cancer. General Nursing, 2025, 23(13): 2548-2553.

[2] He J., Li J., Yin X., Lei Z. Trends in lung cancer incidence in China from 1990 to 2021 and future predictions. Journal of Practical Cardiovascular and Pulmonary Diseases, 2025, 33(8): 1-5.

[3] Hu Y. Investigation of lung cancer incidence in cancer cases in a certain region. Chinese Typical Case Collection, 2025, 19(1): 439-442.

[4] Tian Z., Song Y., Li J., Ma J. Comparison of epidemiological characteristics of lung cancer among adolescents and young adults in China and globally from 1990 to 2021. Chinese Journal of Thoracic and Cardiovascular Surgery, 2025, 32(7): 991-999.

[5] Zheng C., Huang M., Xian G., Zhu Z., Zhou Q., Zhu G., Li B. Research progress on the effects of tobacco and PM2.5 on lung cancer immunotherapy. Biomedical Translation, 2025, 6(2): 80-85.

[6] Kaggle. Lung cancer prediction. 2022-01-01. Available: https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link (Accessed: 2025-07-22).

[7] Long H., Wang X., Li F., Bai Z., Yao M., An Y., Huang Y. Analysis of the global and China-US disease burden and trends of lung cancer attributable to tobacco from 1990 to 2021. Chinese Journal of Thoracic and Cardiovascular Surgery, 2025: 1-8.

[8] Lam V. K., Bentzen S. M., Mohindra P., Nichols E. M., Bhooshan N., Vyfhuis M., Scilla K. A., Feigenberg S. J., Edelman M. J., Feliciano J. L. Obesity is associated with long-term improved survival in definitively treated locally advanced non-small cell lung cancer (NSCLC). Lung Cancer, 2017, 104: 52-57.

[9] Huang L., Zhao M., Mou X., Qu X., Lei J. Research progress on the correlation between lung cancer and obesity. Modern Oncology Medicine, 2025, 33(6): 1034-1038.

[10] Liu Y., Cui Z. Analysis of the disease burden of lung cancer attributable to air pollution in China from 1990 to 2021. Chinese Journal of Health Statistics, 2025, 42(3): 350-354.