Research Article

Combined Influence of Multiple Factors on Lung Cancer and Anxiety: A Probit Model Analysis

Cecilia Zheng

Diamond Bar High School, 21400 Pathfinder Rd, Diamond Bar, CA 91765, USA

ABSTRACT

Lung cancer remains a leading cause of cancer-related mortality worldwide, significantly impacting public health. Concurrently, anxiety is a prevalent psychological disorder known to influence the development and progression of various cancers. This research paper aims to examine the combined influence of various factors, including demographic characteristics, environmental exposures, and physical conditioning data, on the incidence of lung cancer and anxiety. Given the binary nature of both lung cancer and anxiety outcomes, the analyses will first employ separate binary probit regression models to identify significant predictors for each condition independently. Then, joint modeling techniques will be implemented with dual purposes. First, by comparing the results of the individual and joint models, the reliability and robustness of the findings will be enhanced through cross-validation. Second, joint models will enable an investigation into potential endogeneity between lung cancer and anxiety. Addressing this endogeneity is crucial as it can potentially improve model robustness and provide deeper insights into the interrelationship between these two health outcomes. For the study purpose, the demographic factors considered include age and gender. Environmental factors encompass smoking history, alcohol consumption, and peer pressure. Physical conditioning data includes pre-existing health conditions such as yellow fingers, anxiety, chronic disease, fatigue, allergy, wheezing, coughing, shortness of breath, swallowing difficulty, and chest pain. By leveraging advanced statistical modeling techniques, this research seeks to uncover nuanced relationships and potential causal pathways that may exist between lung cancer and anxiety. The findings from this study will contribute to the existing body of knowledge by providing additional case study showing the relationship among a multitude of factors on lung cancer and anxiety, respectively. Also, the endogeneity checking among lung cancer and anxiety may enhance the efficiency of models done by others in the future. The research paper's analysis on lung cancer incidences reveals that age and wheezing are significant predictors. In examining anxiety levels, smoking emerged as a significant predictor, indicating a higher likelihood of anxiety among smokers. In addition, the joint probit models confirmed these findings, with age and wheezing significantly predicting lung cancer incidence, and smoking significantly predicting anxiety levels. No significant endogeneity was observed between lung cancer and anxiety, suggesting that these health outcomes are influenced by different sets of factors. These findings underscore the importance of considering demographic, environmental, and physical conditioning data in understanding and addressing lung cancer and anxiety, and checking their potential endogeneity.

Keywords: Lung Cancer, Anxiety, Binary Probit Model, Joint Models, Endogeneity

Corresponding author: Cecilia Zheng, E-mail: cicizhengg@gmail.com.

Received August 10, 2024**; Accepted** August 25, 2024
https://doi.org/10.70251/HYJR2348.23112

INTRODUCTION

Lung cancer is one of the most common and deadliest cancers worldwide. According to the World Health Organization (WHO), it accounts for approximately 11.6% of all cancer cases and is the leading cause of cancer-related deaths globally, responsible for 18.4% of all cancer deaths (1). In 2020, there were an estimated 2.2 million new cases of lung cancer and 1.8 million deaths (2). The high mortality rate is primarily due to the late-stage diagnosis in many patients, as early-stage lung cancer is often asymptomatic. The burden of lung cancer on public health is substantial, not only due to the high mortality rates but also because of the significant economic costs associated with its treatment and management.

Due to the importance of lung cancer, many strategies have been proposed to address lung cancer issues. These strategies include early detection and screening programs, public health campaigns to reduce smoking rates, improving air quality, promoting healthy lifestyles, and advancing medical treatments through research and clinical trials. Among them, one popular method is to identify the relationship between contributing factors and lung cancer via data analysis. Various research studies have demonstrated this approach, using different models and identifying influential factors. For instance, logistic regression models have been widely used to evaluate the impact of smoking, age, and family history on lung cancer risk (3). Cox proportional hazards models have also been applied to study the influence of occupational exposure and environmental pollutants (4). Machine learning techniques like random forests and neural networks have been employed to explore complex interactions between genetic predispositions and lifestyle factors (5). However, most of the previous papers focus on the impact of lung cancer from a certain type of factor, and very few studies have used a wide range of data covering different fields, which is believed to have combined effects on lung cancer. To fill this gap, a comprehensive dataset from Kaggle that includes various aspects such as demographic information (age, gender, socioeconomic status), environmental exposures (pollutants, smoking history, occupational hazards), and physical conditioning data (pre-existing health conditions). This comprehensive information allows for a more holistic understanding of the multiple factors influencing lung cancer, which could then provide more robust and reliable findings that can inform effective prevention and intervention strategies.

Additionally, anxiety has also been identified as being associated with various cancer incidences. In 1989, Holland studied the impact of cancers on the level of psychological distress experienced by patients, families, and close family members (6). The author illustrated that cancer diagnosis and treatment can lead to significant psychological distress, including anxiety and depression, which varies among individuals based on their personal and social contexts. Further research has demonstrated that anxiety is a prevalent issue among cancer patients, with factors like disease stage, treatment type, and personal coping mechanisms influencing anxiety levels (7). These studies highlight the importance of addressing psychological distress in cancer care to improve overall patient outcomes. To take advantage of the availability of the comprehensive dataset, this paper also aims to explore the relationship between anxiety and lung cancer, along with other pertinent factors. Due to the dichotomous nature of anxiety and lung cancer, two isolated binary probit models were first developed for anxiety and lung cancer, respectively. Past studies have demonstrated many types of statistical models for binary outcomes, such as logistic regression, probit models, and complementary log-log models. The paper chose the probit model mainly due to its ability to handle the cumulative distribution function of the standard normal distribution, which often provides better estimates for probabilities in cases where the outcome distribution does not strictly follow a logistic curve (8). Additionally, joint models were also developed with two primary objectives. First, they provide an opportunity to validate the findings by comparing results from individual and combined models, thus enhancing the robustness and reliability of the results. Second, joint models allow for the investigation of potential endogeneity between lung cancer and anxiety. By addressing endogeneity, the models can possibly correct for potential biases that may arise from unobserved confounders influencing both outcomes simultaneously. This two-stage modeling approach aims to uncover nuanced relationships and potential causal pathways, providing deeper insights into the factors influencing lung cancer and anxiety, ultimately guiding more effective prevention and intervention strategies.

Lung cancer patients often require extensive healthcare resources, including surgery, chemotherapy, radiation therapy, and palliative care. The disease also impacts patients' quality of life, leading to substantial physical, emotional, and financial strain on patients and their families (9). Early detection of lung cancer significantly improves prognosis and survival rates (10).

Low-dose computed tomography (LDCT) screening has been shown to reduce lung cancer mortality by detecting tumors at an earlier, more treatable stage (11). Understanding and addressing risk factors such as smoking, environmental exposures, and genetic predispositions are crucial for developing effective prevention strategies and reducing the overall incidence of lung cancer. Psychological factors, particularly anxiety, have been shown to influence cancer development and progression (12). Anxiety can affect an individual's immune function, hormone levels, and overall health behaviors, potentially creating a conducive environment for cancer cells to thrive (13). Chronic anxiety and stress might lead to changes in cellular processes and immune responses that could promote tumor growth and metastasis. External factors like smoking are well-known primary risk factors for lung cancer. Smoking alone is responsible for about 85% of all lung cancer cases (14). Peer pressure, especially among adolescents and young adults, can lead to the initiation of smoking habits. Chronic diseases such as Chronic Obstructive Pulmonary Disease (COPD) also increase the risk of developing lung cancer (15). The interaction between these external factors and psychological stressors can compound the risk, necessitating a comprehensive approach to prevention and treatment. Numerous studies have focused on developing predictive models for lung cancer, with logistic regression (16) and probit models (17) being a commonly used method. These models typically incorporate variables such as age, smoking history, genetic predisposition, and exposure to environmental toxins. For instance, studies have shown that logistic regression models can effectively predict lung cancer risk based on smoking intensity and duration, as well as occupational exposures. However, there are notable gaps in the current literature. Many models fail to integrate psychological factors such as anxiety and stress, which can also play a crucial role in cancer development and progression. The absence of joint modeling of both psychological and external factors limits the comprehensiveness and accuracy of these predictive models. The need for integrated models that consider both physical and psychological predictors of lung cancer is evident. Integrating anxiety into lung cancer prediction models could provide a more holistic understanding of cancer risk and improve early detection and intervention strategies. To this end, this research aims to fill the gap by studying the combined impact of psychological and external factors on lung cancer risk. This research paper will explore the joint modeling of psychological and external factors in predicting lung cancer risk, addressing an important gap in the and potentially improving outcomes through more comprehensive risk assessment and early intervention strategies.

DATA DESCRIPTION

The data set used in this study is from Kaggle, a well-known platform for data science and machine learning competitions. The specific dataset is titled “Lung Cancer Dataset.” (18) The dataset contains a comprehensive collection of attributes pertinent to individuals' health profiles and potential risk factors for lung cancer. It comprises features such as age, gender, smoking habits, presence of yellow fingers, anxiety levels, peer pressure, chronic diseases, fatigue, allergies, wheezing, alcohol consumption, coughing, shortness of breath, swallowing difficulty, and chest pain. Each record in the dataset provides detailed information that can be utilized to analyze and identify patterns and correlations related to lung cancer risk factors. However, it is acknowledged that there are some limitations to this data, such as the relatively small sample size (3000) which could lead to inaccurate results. This could be easily improved by gathering a larger sample size to improve the validity of the results. (Table 1)

METHODOLOGY

Given the complexity of analyzing data from a large number of variables relying on a single technique or tool often falls short in providing comprehensive insights. Consequently, this research employs a highly valuable analytical method to ensure a thorough exploration of the data. Both isolated and joint probit regression models are selected for their unique strengths in uncovering different aspects of the data and the combined ability to offer a robust analytical framework.

Binary Probit Regression

The probit model is a statistical technique used to model binary outcome variables. It is particularly useful when the dependent variable is binary, meaning it has two possible outcomes, such as the presence or absence of a disease. The probit model links the probability of the occurrence of an event to a set of predictor variables through the cumulative distribution function (CDF) of the standard normal distribution. It is essential in fields like medical research because it provides a way

Table 1. Descriptive Statistics of Variables

Variables	Data Type	Definition	Descriptive Statistics
Gender	Categorical	The biological sex of the individual, often impacting the incidence and type of lung cancer due to differences in smoking rates and hormonal influences. (M for male and F for female.)	M: 1500 (50%); F: 1500 (50%).
Age	Numerical	The age of the individual, with older age being a significant risk factor for lung cancer.	Mean: 55.1; SD: 14.7; Min: 30.0; Max: 80.0.
Smoking	Categorical	The act of inhaling tobacco smoke, a major cause of lung cancer due to carcinogens. (1 for yes, 0 for no)	1: 1,527; 0: 1,473.
Yellow Fingers	Categorical	Discoloration of fingers from smoking, indicating prolonged exposure to tobacco. (1 for yes, 0 for no)	1: 1,458; 0: 1,542.
Anxiety	Categorical	Psychological condition that can influence smoking habits and overall health. (1 for yes, 0 for no)	1: 1,518; 0: 1,482.
Peer Pressure	Categorical	Social influence that can lead to smoking and other risky behaviors associated with lung cancer. (1 for yes, 0 for no)	1: 1,503; 0: 1,497.
Chronic Disease	Categorical	Long-term health conditions that can weaken the body's defense mechanisms and potentially increase cancer risk. (1 for yes, 0 for no)	1: 1,471; 0: 1,529.
Fatigue	Categorical	A symptom of chronic illness or lung cancer, reflecting the body's decreased energy levels. (1 for yes, 0 for no)	1: 1,531; 0: 1,469.
Allergy	Categorical	Immune responses that may affect respiratory health, though not directly linked to lung cancer. (1 for yes, 0 for no)	1: 1,480; 0: 1,520.
Wheezing	Categorical	A high-pitched sound during breathing, often related to respiratory issues including lung cancer. (1 for yes, 0 for no)	1: 1,508; 0: 1,492.
Alcohol Consumption	Categorical	Intake of alcohol, which can contribute to overall cancer risk through immune suppression. (1 for yes, 0 for no)	1: 1,526; 0: 1,474.
Coughing	Categorical	A common symptom of lung cancer, particularly persistent or with blood. (1 for yes, 0 for no)	1: 1,468; 0: 1,532.
Shortness of Breath	Categorical	Difficulty in breathing, a significant symptom of lung cancer. (1 for yes, 0 for no)	1: 1,536; 0: 1,464.
Swallowing Difficulty	Categorical	Trouble swallowing, which can be associated with advanced stages of lung cancer. (1 for yes, 0 for no)	1: 1,531; 0: 1,469
Chest Pain	Categorical	Discomfort in the chest area, a symptom often associated with lung cancer. (1 for yes, 0 for no)	1: 1,504; 0: 1,496.
Lung Cancer	Categorical	Diagnoses of whether the subject has lung cancer.	Yes: 1518 No: 1482

to estimate the likelihood of binary outcomes based on various predictors. It transforms the linear combination of predictor variables into a probability that lies between 0 and 1, ensuring a realistic modeling of binary events. This makes it especially suitable for scenarios where the response variable is not continuous but categorical. In this research, the probit model is applied to investigate the incidence of lung cancer. The presence of lung cancer is modeled as a binary outcome, with anxiety levels and other relevant covariates as predictor variables. By using the probit model, researchers can estimate the probability that an individual has lung cancer based on their anxiety levels and other factors. This is crucial for understanding the relationship between mental health and physical health outcomes.

Binary Probit Model for Lung Cancer Incidences. The first model is a binary probit model for lung cancer incidences. In this model, the presence of lung cancer is the dependent variable, and anxiety levels, along with many other covariates, are included. The advantage of using a binary probit model is that it is suitable for binary outcome variables.

The latent variable Y^* is defined as (19):

$$Y_{1i}^* = x_i \beta_1 + \epsilon_i \tag{1}$$

Where Y is the latent variable representing the propensity to have lung cancer, x_i represents the covariates (including anxiety levels), β_1 is the vector of coefficients, and ϵ_i is the error term.

The probability of having lung cancer can be calculated by (19):

$$P\,(Y_{1i} = 1) = \Phi\,(x_i \beta_1) \tag{2}$$

Where Φ denotes the cumulative distribution function (CDF) of the standard normal distribution.

Binary Probit Model for Anxiety Levels. The second model is a binary probit model for high anxiety levels. In such a model, anxiety is the dependent variable, and lung cancer incidences, among other variables, are covariates. The choice of the model rests on the following assumptions:

a) anxiety follows a binary distribution (high vs. low);

b) the errors are uncorrelated and have homogeneous variance. The model can be written as follows:

$$Y_{2i}^* = x_i \beta_2 + \epsilon_i \tag{3}$$

where Y_{2i}^* is the latent variable representing the propensity to have high anxiety, x_1 represents the covariates (including lung cancer incidences), β_2 is the vector of coefficients, and ϵ_i is the error term.

The probability of high anxiety can be calculated by:

$$P\,(Y_{2i} = 1) = \Phi\,(x_i \beta_2) \tag{4}$$

Joint Binary Probit Model for Lung Cancer and Anxiety Levels. The third model is a joint binary probit model for lung cancer and high anxiety levels. In this model, the presence of lung cancer and high anxiety are both dependent variables, and they are modeled together to account for their potential relationship.

In the joint model, we assume that the error terms ϵ_{1i} and ϵ_{2i} may be correlated. This can be captured using a bivariate normal distribution for the error terms:

$$(\epsilon_{1i}, \epsilon_{2i}) \sim \mathrm{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & p\\ p & 1\end{pmatrix}\right) \tag{5}$$

where p is the correlation between the error terms of the two equations.

To estimate the parameters, we maximize the joint likelihood function, which takes into account the correlation between the two outcomes.

Modeling Evaluation: Log Likelihood. The log likelihood function provides a measure of how well the statistical model fits the observed data. It quantifies the likelihood of the observed data under the assumptions of the model.

The likelihood function L(θ) is the joint probability of the observed data as a function of the parameters θ.

For example, for a set of n independent and identically distributed (i.i.d.) observations $x_1, x_2, \ldots, x_n$, the likelihood function is: (20)

$$L(\theta) = \prod_{i=1}^{n} f(x_i; \theta) \tag{6}$$

where $f\,(x_i;\ \theta)$ is the probability density function (pdf) for continuous variables or the probability mass function (pmf) for discrete variables, parameterized by θ.

The log likelihood is the natural logarithm of the likelihood function, which simplifies the multiplication into addition.

$$Log\ L(\theta) = Log\ (\prod_{i=1}^{n} f(x_i; \theta)) \tag{7}$$

The log of a product is the sum of the logs.

$$Log\ L(\theta) = Log\ (\sum_{\{i=1\}}^{\{n\}} Log\ f(x_i; \theta)) \tag{8}$$

RESULTS AND CONCLUSIONS

Binary Probit Model for Lung Cancer Incidences Results

As shown in Table 2, In a probit regression analysis examining factors associated with lung cancer, several key variables yielded surprising results. Age demonstrated a statistically significant negative association with lung cancer (coefficient = -0.0839, p = 0.007), suggesting that as age increases, the likelihood of lung cancer slightly decreases, which contradicts common medical understanding. Conversely, yellow fingers, typically an indicator of heavy smoking, showed a positive but non-significant association (coefficient = 0.0910, p = 0.239), which is unexpected given the established link between smoking and lung cancer. Wheezing, another factor analyzed, exhibited a statistically significant negative association with lung cancer (coefficient = -0.1413, p = 0.010), indicating that individuals with wheezing are less likely to have lung cancer, a result that diverges from typical clinical expectations. These findings highlight potential complexities in lung cancer risk factors and suggest the need for further investigation to validate these associations. Chronic disease and fatigue were both analyzed and neither showed significant associations with lung cancer, with coefficients of 0.0109 (p = 0.843) and 0.0019 (p = 0.973), respectively. Allergy had a negative, non-significant association (coefficient = -0.0114, p = 0.836). Wheezing exhibited a statistically significant negative association with lung cancer (coefficient = -0.1413, p = 0.010), indicating that individuals with wheezing are less likely to have lung cancer, which diverges from typical clinical expectations. Alcohol consumption showed a negative but non-significant association (coefficient = -0.0942, p = 0.087). Other factors such as coughing (coefficient = 0.0317, p = 0.565), shortness of breath (coefficient = -0.0691, p = 0.209), swallowing difficulty (coefficient = -0.0155, p = 0.779), and chest pain (coefficient = 0.0355, p = 0.519) did not show significant associations with lung cancer. Gender (M) also showed a non-significant negative association (coefficient = -0.0794, p = 0.148).

The analysis identified several predictors with both positive and negative associations with lung cancer incidence. Positive coefficients were observed for smoking (0.0190), yellow fingers (0.0649), chronic

Table 2. Modeling Result for Binary Probit of Lung Cancer

Variable	Coefficient	Std Error	z	P > |z|
Age	**-0.0039**	0.002	-2.094	0.036
Smoking	0.0190	0.055	0.344	0.731
Yellow Fingers	0.0649	0.055	1.178	0.239
Anxiety	-0.0364	0.055	-0.662	0.508
Peer Pressure	-0.0384	0.055	-0.698	0.485
Chronic Disease	0.0109	0.055	0.198	0.843
Fatigue	0.0019	0.055	0.034	0.973
Allergy	-0.0114	0.055	-0.207	0.836
Wheezing	**-0.1413**	0.055	-2.569	0.010
Alcohol Consumption	-0.0942	0.055	-1.710	0.087
Coughing	0.0317	0.055	0.575	0.565
Shortness of Breath	-0.0691	0.055	-1.255	0.209
Swallowing Difficulty	-0.0155	0.055	-0.281	0.779
Chest Pain	0.0355	0.055	0.645	0.519
Gender (M)	-0.0794	0.055	-1.445	0.148
Modeling Evaluation	Log-Likelihood: -1445.2			

Note: Bond fonts indicated the significance level of 0.05.

disease (0.0109), fatigue (0.0019), coughing (0.0317), and chest pain (0.0355). Conversely, negative coefficients were found for age (-0.0039), anxiety (-0.0364), peer pressure (-0.0384), allergy (-0.0114), wheezing (-0.1413), alcohol consumption (-0.0942), shortness of breath (-0.0691), swallowing difficulty (-0.0155), and gender (M) (-0.0794). Notably, age demonstrated a statistically significant negative association with lung cancer ($p = 0.036$), suggesting that as age increases, the likelihood of lung cancer slightly decreases, which contradicts common medical understanding. Wheezing also showed a significant negative association with lung cancer ($p = 0.010$), indicating that individuals with wheezing are less likely to have lung cancer, diverging from typical clinical expectations. Other factors, despite showing positive or negative associations, did not reach statistical significance. These findings highlight the complexities in lung cancer risk factors and suggest the need for further research to validate these associations and understand the underlying mechanisms. The counterintuitive results, such as the negative association of age and wheezing with lung cancer, underscore the importance of comprehensive data analysis in revealing nuanced health relationships.

The log-likelihood value for Table 2 is -1445.2, indicating the fit of the model to the observed data. In the context of statistical modeling, the log-likelihood function measures how well the model's parameters explain the observed outcomes. A higher (less negative) log-likelihood value suggests a better fit, as it implies that the model's predictions are more consistent with the actual data. Conversely, a lower (more negative) log-likelihood value, like -1445.2, suggests that there might be room for improvement in the model's specification or that certain important variables or interactions may be missing. In this specific model, the log-likelihood value reflects the combined effect of all the variables included in predicting lung cancer incidences.

Binary Probit Model for Anxiety Levels

As revealed in Table 3, In a probit regression analysis examining factors associated with anxiety, smoking was found to have a statistically significant positive association with anxiety (coefficient = 0.0985, $p = 0.046$), indicating that individuals who smoke are more likely to experience anxiety. Conversely, while allergy also showed a positive association with anxiety (coefficient =

Table 3. Modeling Result for Binary Probit of Anxiety

Variable	Coefficient	Std Error	z	P > |z|
Age	-0.0013	0.002	-0.718	0.473
Smoking	**0.1096**	0.055	1.991	0.046
Yellow Fingers	-0.0548	0.055	-0.996	0.319
Peer Pressure	0.0097	0.055	0.177	0.860
Chronic Disease	-0.0330	0.055	-0.599	0.549
Fatigue	0.0199	0.055	0.362	0.717
Allergy	0.0873	0.055	1.587	0.113
Wheezing	0.0191	0.055	0.346	0.729
Alcohol Consumption	0.0433	0.055	0.787	0.431
Coughing	0.0395	0.055	0.719	0.472
Shortness of Breath	0.0591	0.055	1.075	0.282
Swallowing Difficulty	-0.0438	0.055	-0.796	0.426
Chest Pain	-0.0576	0.055	-1.048	0.295
Gender (M)	0.0607	0.055	1.104	0.270
Lung Cancer (Yes)	0.0366	0.055	0.665	0.506
Modeling Evaluation	Log-Likelihood: -1448.4			

Note: Bond fonts indicated the significance level of 0.05.

0.1033), this relationship was not statistically significant (p = 0.113). These findings suggest that smoking is a significant predictor of anxiety, whereas the association between allergies and anxiety, although present, requires further investigation to establish statistical significance. However, other factors yielded unexpected results. For instance, yellow fingers, typically an indicator of heavy smoking, showed a negative but non-significant association with anxiety (coefficient = -0.0548, p = 0.319), which contradicts the expectation that heavy smoking indicators would correlate with higher anxiety. Similarly, peer pressure showed a positive but non-significant association with anxiety (coefficient = 0.0097, p = 0.860), suggesting that peer pressure may not be as influential in anxiety levels as previously thought. Chronic disease, another factor typically associated with increased anxiety due to the stress of managing long-term health conditions, showed a negative but non-significant association with anxiety (coefficient = -0.0330, p = 0.549). Additionally, shortness of breath had a positive but non-significant association with anxiety (coefficient = 0.0591, p = 0.282), which is surprising given that shortness of breath is often a symptom associated with anxiety disorders.

The probit regression analysis for anxiety identified several predictors with both positive and negative coefficients. Predictors with positive coefficients include smoking (0.1096), peer pressure (0.0097), fatigue (0.0199), allergy (0.0873), wheezing (0.0191), alcohol consumption (0.0433), coughing (0.0395), shortness of breath (0.0591), gender (M) (0.0607), and lung cancer (yes) (0.0366). Conversely, predictors with negative coefficients include age (-0.0013), yellow fingers (-0.0548), chronic disease (-0.0330), swallowing difficulty (-0.0438), and chest pain (-0.0576). Among these, smoking was the only variable with a statistically significant positive association with anxiety (p = 0.046), indicating that smokers are more likely to experience anxiety. Surprisingly, variables such as yellow fingers and chronic disease, typically associated with higher anxiety levels, showed negative but non-significant associations. These findings suggest that while smoking is a significant predictor of anxiety, the relationships between other factors and anxiety are more complex and may involve underlying mechanisms that require further investigation. The analysis underscores the importance of addressing smoking as a modifiable risk factor for anxiety and highlights the need for a more nuanced understanding of how various factors contribute to anxiety.

The log-likelihood value for Table 3 is -1448.4, indicating the fit of this model to the observed data related to anxiety. Similar to the Table 2, the log-likelihood value measures how well the parameters of this model explain the observed outcomes. In this case, a value of -1448.4 suggests the model's fit to the data is relatively similar to the fit of the lung cancer model, though slightly worse (as indicated by a more negative value). This value implies that the model could potentially be improved by including additional relevant variables or refining the existing ones to better capture the factors influencing anxiety. Overall, the log-likelihood provides a numerical summary of the model's effectiveness in predicting anxiety based on the included variables.

Joint Binary Probit Models for Lung Cancer and Anxiety Levels

In review of Table 4, the Lung Cancer Model, the age variable has a coefficient of -0.0839 and a p-value of 0.035. The negative coefficient suggests that as age increases, the likelihood of developing lung cancer slightly decreases. This finding is counterintuitive as older age is generally considered a risk factor for many cancers, including lung cancer. This result might indicate potential issues with the data or model specification. One possible explanation is that the data set might not adequately represent the age distribution typically seen in lung cancer patients, or there might be confounding variables not accounted for in the model. The wheezing variable has a coefficient of -0.1416 and a p-value of 0.010. It is associated with a lower likelihood of lung cancer. Other factors yielded unexpected results as well. For instance, smoking, a well-known risk factor for lung cancer, showed a positive but non-significant association (coefficient = 0.0205, p = 0.709). Yellow fingers, typically an indicator of heavy smoking, also showed a positive but non-significant association with lung cancer (coefficient = 0.0641, p = 0.244), which contradicts the expectation that heavy smoking indicators would correlate with higher lung cancer risk. Peer pressure showed a negative but non-significant association with lung cancer (coefficient = -0.0383, p = 0.487), suggesting that peer pressure may not be as influential in lung cancer risk as previously thought. Chronic disease, another factor typically associated with increased lung cancer risk due to the stress of managing long-term health conditions, showed a positive but non-significant association with lung cancer (coefficient = 0.0105, p = 0.849). Additionally, shortness of breath had a negative but non-significant association with lung cancer (coefficient = -0.0682, p = 0.215), which is surprising given that shortness of breath is often a symptom associated with lung conditions, including lung cancer. This result

is surprising because wheezing is commonly associated with respiratory conditions that can be related to lung cancer. The unexpected negative association might be due to wheezing being more commonly reported or diagnosed in individuals with non-cancerous respiratory conditions, leading to a misleading inverse relationship in this model. Other variables such as smoking, yellow fingers, peer pressure, chronic disease, fatigue, allergy, alcohol consumption, coughing, shortness of breath, swallowing difficulty, chest pain, and gender were not significant predictors of lung cancer in this analysis. This is particularly surprising for smoking, a well-established risk factor for lung cancer. The lack of significance could be due to several factors such as data quality and multicollinearity. For example, there might be issues with how smoking status or intensity was measured and recorded. In addition, high correlation among predictors could inflate standard errors, making it difficult to detect significant associations.

In Table 4, the log-likelihood value for the Joint Binary Probit Model for Lung Cancer is -1445.4. This value reflects the fit of the model to the observed data, considering multiple variables simultaneously to predict lung cancer incidences. A log-likelihood of -1445.4 suggests that the model has a comparable fit to the previously presented models, specifically focusing on lung cancer. The relatively close values of -1445.2 and -1445.4 between this and the earlier model indicate that there might not be significant differences in their ability to explain the data, though slight variations could exist due to the different variables considered.

Modeling Results Comparison

In Table 5, the Anxiety model, the smoking variable has a coefficient of 0.1039 and a p-value of 0.046. The positive coefficient indicates that individuals who smoke are more likely to experience anxiety. This aligns with existing literature, which suggests that smoking can be both a coping mechanism for anxiety and a factor that exacerbates it. The nicotine in cigarettes may temporarily relieve anxiety symptoms, but over time, dependence and withdrawal can increase anxiety levels. Other variables such as age, yellow fingers, peer pressure, chronic disease, fatigue, allergy, wheezing, alcohol consumption, coughing, shortness of breath, swallowing difficulty, chest pain, and gender did not show significant associations with anxiety. For instance, yellow fingers, typically an indicator of heavy smoking, showed a negative but

Table 4. Joint Modeling Result for Binary Probit of Lung Cancer

Variable	Coefficient	Std Error	z	P > |z|
Age	**-0.0039**	0.002	-2.105	0.035
Smoking	0.0205	0.055	0.373	0.709
Yellow Fingers	0.0641	0.055	1.165	0.244
Peer Pressure	-0.0383	0.055	-0.695	0.487
Chronic Disease	0.0105	0.055	0.190	0.849
Fatigue	0.0022	0.055	0.039	0.969
Allergy	-0.0102	0.055	-0.185	0.853
Wheezing	**-0.1410**	0.055	-2.564	0.010
Alcohol Consumption	-0.0935	0.055	-1.699	0.089
Coughing	0.0323	0.055	0.586	0.558
Shortness of Breath	-0.0682	0.055	-1.240	0.215
Swallowing Difficulty	-0.0161	0.055	-0.293	0.769
Chest Pain	0.0347	0.055	0.630	0.528
Gender (M)	-0.0786	0.055	-1.430	0.153
Modeling Evaluation	Log-Likelihood: -1445.4			

Note: Bond fonts indicated the significance level of 0.05.

non-significant association with anxiety (coefficient = -0.0539, p = 0.327), which contradicts the expectation that heavy smoking indicators would correlate with higher anxiety. Similarly, peer pressure showed a positive but non-significant association with anxiety (coefficient = 0.0092, p = 0.867), suggesting that peer pressure may not be as influential in anxiety levels as previously thought. Chronic disease, another factor typically associated with increased anxiety due to the stress of managing long-term health conditions, showed a negative but non-significant association with anxiety (coefficient = -0.0328, p = 0.551). Additionally, shortness of breath had a positive but non-significant association with anxiety (coefficient = 0.0581, p = 0.290), which is surprising given that shortness of breath is often a symptom associated with anxiety disorders. The non-significance of these variables could be caused by the complex nature of anxiety since anxiety disorders are influenced by a multitude of factors, including genetics, environment, and psychological factors, which might not be fully captured by the variables included in this model. Another reason could be a problem with the measurement as the data might not accurately capture all dimensions of anxiety.

In Table 5, the log-likelihood value for the Joint Binary Probit Model for Anxiety is -1448.6. A log-likelihood of -1448.6 suggests that the model's fit to the data is slightly less effective compared to the lung cancer models previously discussed. This value reflects the model's ability to explain the variability in anxiety outcomes based on the included predictors. While the fit is not exceptionally strong, it provides a quantitative measure of the model's overall performance, indicating that further refinement or inclusion of additional relevant variables might improve the model's explanatory power.

The findings suggest that while certain factors like smoking are clearly linked to anxiety, and age and wheezing to lung cancer, there is a need for more comprehensive models that include a wider range of predictors and potentially interactions between them. The unexpected results for age and wheezing in the lung cancer model indicate the need for scrutiny of the data and model specification. Future research should aim to incorporate a broader set of variables, improve data quality, and explore more sophisticated modeling techniques to better understand the determinants of lung cancer and anxiety.

CONCLUSION

In the lung cancer model, the significant factors included age and wheezing. Age showed a statistically significant

Table 5. Joint Modeling Result for Binary Probit of Anxiety

Variable	Coefficient	Std Error	z	P > |z|
Age	-0.0014	0.002	-0.749	0.454
Smoking	**0.1099**	0.055	1.996	0.046
Yellow Fingers	-0.0539	0.055	-0.979	0.327
Peer Pressure	0.0092	0.055	0.167	0.867
Chronic Disease	-0.0328	0.055	-0.596	0.551
Fatigue	0.0199	0.055	0.363	0.717
Allergy	0.0871	0.055	1.584	0.113
Wheezing	0.0170	0.055	0.310	0.757
Alcohol Consumption	0.0420	0.055	0.763	0.445
Coughing	0.0400	0.055	0.726	0.468
Shortness of Breath	0.0581	0.055	1.058	0.290
Swallowing Difficulty	-0.0440	0.055	-0.800	0.424
Chest Pain	-0.0571	0.055	-1.038	0.299
Gender (M)	0.0595	0.055	1.083	0.279
Modeling Evaluation	Log-Likelihood: -1448.6			

Note: Bond fonts indicated the significance level of 0.05

negative association with lung cancer (coefficient = -0.0039, p = 0.035), which is counterintuitive since older age is generally considered a risk factor for cancer. Wheezing also exhibited a significant negative association with lung cancer (coefficient = -0.1410, p = 0.010), suggesting that individuals with wheezing are less likely to develop lung cancer. This is surprising because wheezing is commonly associated with respiratory conditions related to lung cancer. Unexpectedly, smoking, a well-known risk factor for lung cancer, showed a positive but non-significant association (coefficient = 0.0205, p = 0.709). Similarly, yellow fingers, another indicator of heavy smoking, showed a positive but non-significant association with lung cancer (coefficient = 0.0641, p = 0.244). Other non-significant predictors included peer pressure (coefficient = -0.0383, p = 0.487), chronic disease (coefficient = 0.0105, p = 0.849), fatigue (coefficient = 0.0022, p = 0.969), allergy (coefficient = -0.0102, p = 0.853), alcohol consumption (coefficient = -0.0935, p = 0.089), coughing (coefficient = 0.0323, p = 0.558), shortness of breath (coefficient = -0.0682, p = 0.215), swallowing difficulty (coefficient = -0.0161, p = 0.769), chest pain (coefficient = 0.0347, p = 0.528), and gender (M) (coefficient = -0.0786, p = 0.153). The non-significance of smoking is particularly surprising and may be due to data quality issues, measurement errors, or multicollinearity among predictors.

In the anxiety model, smoking was the only significant factor (coefficient = 0.1099, p = 0.046), indicating that individuals who smoke are more likely to experience anxiety. This aligns with existing literature that suggests smoking can exacerbate anxiety over time. Other variables such as age (coefficient = -0.0014, p = 0.454), yellow fingers (coefficient = -0.0539, p = 0.327), peer pressure (coefficient = 0.0092, p = 0.867), chronic disease (coefficient = -0.0328, p = 0.551), fatigue (coefficient = 0.0199, p = 0.717), allergy (coefficient = 0.0871, p = 0.113), wheezing (coefficient = 0.0170, p = 0.757), alcohol consumption (coefficient = 0.0420, p = 0.445), coughing (coefficient = 0.0400, p = 0.468), shortness of breath (coefficient = 0.0581, p = 0.290), swallowing difficulty (coefficient = -0.0440, p = 0.424), chest pain (coefficient = -0.0571, p = 0.299), and gender (M) (coefficient = 0.0595, p = 0.279) did not show significant associations with anxiety.

The analysis revealed no significant endogeneity between lung cancer and anxiety. Despite the initial hypothesis that these two conditions might be interrelated, the results showed that their predictors did not significantly overlap, and the conditions did not influence each other in a detectable way within this dataset. The joint modeling techniques used did not reveal any underlying endogeneity, suggesting that lung cancer and anxiety are influenced by different sets of factors. This absence of endogeneity indicates that lung cancer and anxiety should be studied independently rather than assuming a direct causal relationship between the two.

The findings suggest that while smoking is clearly linked to anxiety, and age and wheezing to lung cancer, other factors showed surprising non-significant associations. The lack of significance for well-established risk factors like smoking in the lung cancer model and the unexpected negative associations for age and wheezing highlight potential issues with data quality, measurement, and model specification. This indicates the need for more comprehensive models that include a wider range of predictors and potentially interactions between them to fully understand the determinants of lung cancer and anxiety. Addressing endogeneity between lung cancer and anxiety was crucial for this study, but the absence of significant overlap or causal influence between these conditions suggests that they can be studied independently. This finding underscores the complexity of these health outcomes and the importance of targeted research to address the specific risk factors associated with each condition.

In conclusion, while the analysis confirmed some expected relationships, it also revealed complexities and unexpected results that warrant further investigation. Future research should aim to incorporate a broader set of variables, improve data quality, and explore more sophisticated modeling techniques to better understand the risk factors for lung cancer and anxiety. This comprehensive approach will enhance the reliability and validity of findings, ultimately guiding more effective prevention and intervention strategies.

ACKNOWLEDGEMENT:

I would like to express my gratitude to the anonymous reviewers and editors for their valuable time and constructive feedback on my research paper. I also extend my thanks to Kaggle for providing the dataset that made this research and analysis possible.

REFERENCE

1. Dakubo GD, Dakubo GD. Global burden of cancer and the call to action. Cancer Biomarkers in Body Fluids: Biomarkers in Proximal Fluids. 2019; 1-20.
2. Ferlay J, Colombet M, Soerjomataram I, Parkin DM, Piñeros M, Znaor A & Bray F. Cancer statistics for the year

2020: An overview. International journal of cancer. 2021; 149 (4): 778-789.
3. Coté ML, Liu M, Bonassi S, Neri M, Schwartz AG, Christiani DC, ... & Hung RJ. Increased risk of lung cancer in individuals with a family history of the disease: a pooled analysis from the International Lung Cancer Consortium. European journal of cancer. 2012; 48 (13): 1957-1968.
4. Bauleo L, Bucci S, Antonucci C, Sozzi R, Davoli M, Forastiere F & Ancona C. Long-term exposure to air pollutants from multiple sources and mortality in an industrial area: a cohort study. Occupational and Environmental Medicine. 2019; 76 (1): 48-57.
5. Koo CL, Liew MJ, Mohamad MS & Mohamed Salleh AH. A Review for Detecting Gene-Gene Interactions Using Machine Learning Methods in Genetic Epidemiology. BioMed research international. 2013; 2013 (1): 432375.
6. Holland JC. Anxiety and cancer: the patient and the family. The Journal of clinical psychiatry. 1989; 50: 20-25.
7. Stark DPH & House A. Anxiety in cancer patients. British journal of cancer. 2000; 83 (10): 1261-1267.
8. Wilks DS. Extending logistic regression to provide full-probability-distribution MOS forecasts. Meteorological Applications: A journal of forecasting, practical applications, training techniques and modelling. 2009; 16 (3): 361-368.
9. Covinsky KE, Goldman L, Cook EF, Oye R, Desbiens N, Reding D, ... & Murphy DJ. The impact of serious illness on patients' families. Jama. 1994; 272 (23): 1839-1844.
10. Blandin Knight S, Crosbie PA, Balata H, Chudziak J, Hussell T & Dive C. Progress and prospects of early detection in lung cancer. Open biology. 2017; 7 (9): 170070.
11. Zhao SJ & Wu N. Early detection of lung cancer: Low-dose computed tomography screening in C hina. Thoracic Cancer. 2015; 6 (4): 385-389.
12. Garssen B. Psychological factors and cancer development: evidence after 30 years of research. Clinical psychology review. 2004; 24 (3): 315-338.
13. Antoni MH. Psychosocial intervention effects on adaptation, disease course and biobehavioral processes in cancer. Brain, behavior, and immunity. 2013; 30: S88-S98.
14. Parkin DM, Pisani P, Lopez AD & Masuyer E. At least one in seven cases of cancer is caused by smoking. Global estimates for 1985. International journal of cancer. 1994; 59 (4): 494-504.
15. Takiguchi Y, Sekine I, Iwasawa S, Kurimoto R & Tatsumi K. Chronic obstructive pulmonary disease as a risk factor for lung cancer. World journal of clinical oncology. 2014; 5 (4): 660.
16. Spitz MR, Hong WK, Amos CI, Wu X, Schabath MB, Dong Q, ... & Etzel CJ. A risk model for prediction of lung cancer. Journal of the National Cancer Institute. 2007; 99 (9): 715-726.
17. Puntoni R, Toninelli F, Zhankui L & Bonassi S. Mathematical modelling in risk/exposure assessment of tobacco related lung cancer. Carcinogenesis. 1995; 16 (7): 1465-1471.
18. Nath A. (n.d.). Lung cancer dataset. Kaggle. Retrieved August 3, 2024.
19. Cakmakyapan S & Goktas A. A comparison of binary logit and probit models with a simulation study. Journal of Social and Economic statistics. 2013; 2 (1): 1-17.
20. Rosseel Y. Evaluating the observed log-likelihood function in two-level structural equation modeling with missing data: From formulas to R code. Psych. 2021; 3 (2): 197-232.