Open Access Library Journal
2024, Volume 11, e11905
ISSN Online: 2333-9721
ISSN Print: 2333-9705

Estimating the Likelihood of Cancer Occurrence Based on Patient Data and Lifestyle Factors: A Comparison between Logit and Probit Regression

Wakeel A. Kasali[1], Barakat O. Ogunjoun[2], Abdullahi O. Olakehinde[1], Emmanuel O. Oderinde[2], Abibat O. Salam[3]

[1]Department of Statistics, Ladoke Akintola University of Technology, Ogbomoso, Nigeria
[2]Department of Statistics, Federal University of Technology Akure, Nigeria
[3]Department of Statistics, University of Ilorin, Kwara, Nigeria
Email: wakasali@student.lautech.edu.ng, ogunjounsta188543@futa.edu.ng, aoolakehinde@student.lautech.edu.ng, oderindesta174941@futa.edu.ng, debolaabdsalam@gmail.com

How to cite this paper: Kasali, W.A., Ogunjoun, B.O., Olakehinde, A.O., Oderinde, E.O. and Salam, A.O. (2024) Estimating the Likelihood of Cancer Occurrence Based on Patient Data and Lifestyle Factors: A Comparison between Logit and Probit Regression. *Open Access Library Journal*, **11**: e11905.
https://doi.org/10.4236/oalib.1111905

Received: July 5, 2024
Accepted: August 26, 2024
Published: August 29, 2024

Abstract

This study compares logit and probit regression models to analyze the likelihood of lung cancer occurrence based on patient data and lifestyle factors. The results reveal significant associations between lung cancer and patients' age, gender, smoking status, yellow finger, anxiety, chronic disease, fatigue, allergy, wheezing, alcohol consumption, coughing, shortness of breath, swallowing difficulty, and chest pain. The result shows that the probit model better predicts lung cancer based on the factors observed as its diagnostic performs better than logit. Recommendations based on the findings emphasize the need to create awareness about the significant factors that influence lung cancer. Conclusively, this study contributes to a deeper understanding of the patient's data and lifestyle factors influencing lung cancer and provides valuable insights.

Subject Areas

Statistics and Public Health

Keywords

Logit, Probit, and Lung Cancer

1. Introduction

The growth of many diseases is associated with the abnormal invasion of cells into

other body parts, contributing to the global health burden. In 2023, the World Health Organization reported over 19.3 million new cases of cancer, resulting in 10 million cancer-related deaths worldwide. Among the various types of cancer, lung cancer stands out as one of the most common and deadliest, with approximately a quarter of all cancer-related deaths in America attributed to lung cancer, according to the American Lung Association (2023). Lung cancer is broadly categorized into small-cell lung carcinoma (SCLC) and non-small cell lung carcinoma (NSCLC), with NSCLC accounting for the majority (80%-85%) of cases [1]. Early detection and risk assessment are crucial in improving treatment outcomes and survival rates [2].

The US Department of Health and Human Services (2014) identifies several prominent risk factors for lung cancer, with tobacco smoke exposure being the leading cause, accounting for nearly 80% of lung cancer deaths. Other significant risk factors include exposure to radon gas, air pollution, asbestos, and occupational hazards, such as diesel exhaust [3] [4]. Improving patient outcomes and treatment strategies requires early detection and risk prediction for lung cancer.

This paper tends to determine the significant predictors of lung cancer using logit and probit models. The paper will compare the results of the logit and probit models and choose the best model for estimating the likelihood of lung cancer.

Probit and Logit regression models are recognized as powerful statistical tools for predicting binary outcomes, including the presence or absence of disease [5]. These models analyze extensive patient data, including lifestyle factors and health details, to predict intervention opportunities and tailor personalized treatment strategies [6]. These models pinpoint important risk variables linked to lung cancer and help create prediction models for improved treatment and outcomes by utilizing lifestyle, medical history, environmental exposures, and demographic traits.

It is reasonable to emphasize that every statistical model has limitation(s) when considered for application, and no model is correct. This is why it's paramount to compare at least two models. After determining the significant predictors of lung cancer using logit and probit models, this study will compare the results of the logit and probit models and choose the best model for estimating the likelihood of lung cancer. Hence to prevent overfitting, Information Criteria (IC) value will be identified from the analysis for selecting between the logit and probit models.

2. Materials and Methods

2.1. Materials

This is an observational study with the data sourced from an open-access Kaggle website [7]. It is a retrospective review with 310 observations of lifestyle factors that may contribute to lung cancer. The variables considered for the study are patient-specific and are age, gender, smoking, yellow finger, anxiety, peer pressure, chronic disease, fatigue, allergy, wheezing, alcohol consumption, coughing, shortness of breath, swallowing difficulty, chest pain, and lung cancer. The response variable is lung cancer status; the other lifestyle factors are the explanatory variables. The dataset contains binary categorical variables, represented as an

integer variable (1 or 2) where one typically indicates that the attribute is absent, and two indicate that it is present [8].

2.2. Methods

Analyzing classification problems, logit and probit regression is applicable to estimate the probability of an outcome occurring based on one or more predictor variables which can either be categorical or continuous [9].

2.2.1. Probit Regression Model for Lung Cancer (LC)

The LC is a binary variable representing (No Cancer = 0, Cancer = 1). The model

$$\text{LC} = \beta_0 + \beta_1 \text{Gender} + \beta_2 \text{Age} + \beta_3 \text{Alcohol} + \cdots + \beta_{15} \text{Chest pain} + \varepsilon \tag{1}$$

$$\begin{aligned} &p\left(\text{Gender}, \text{Age}, \text{Alcohol}, \cdots, \text{Chest pain}\right) \\ &= \phi\left(\beta_0 + \beta_1 \text{Gender} + \beta_2 \text{Age} + \beta_3 \text{Alcohol} + \cdots + \beta_{15} \text{Chest pain} + \varepsilon\right) \end{aligned} \tag{2}$$

is the population probit models with multiple repressors Age, Gender, Alcohol, …, Chest pain and $\phi(\cdot)$ is the cumulative standard normal distribution function.

2.2.2. Logit Regression Model for Lung Cancer (LC)

The population logit regression function is

$$\begin{aligned} &P\left(\text{LC} = 1 \mid \text{Age}, \text{Gender}, \text{Alcohol}, \cdots, \text{Chest pain}\right) \\ &= F\left(\beta_0 + \beta_1 \text{Gender} + \beta_2 \text{Age} + \beta_3 \text{Alcohol} + \cdots + \beta_{15} \text{Chest pain}\right) \\ &= \frac{1}{1 + \ell^{-\left(\beta_0 + \beta_1 \text{Gender} + \beta_2 \text{Age} + \cdots + \beta_{15} \text{Chest pain}\right)}} \end{aligned} \tag{3}$$

The Logit regression is very similar to probit only that a different cumulative distribution function (CDF) is used:

$$F(x) = \frac{1}{1 + \ell^{-x}} \tag{4}$$

is the CDF of standard logistic distribution.

2.2.3. Alkaike Information Criteria (AIC) and Bayesian Information Criteria (BIC) of the Probit and Logit Model

AIC and BIC are model selection criteria. They measure the goodness of fits of models. Lower AIC or BIC indicates the preferred model**.** The empirical formula for calculating AIC and BIC is given as:

$$\text{AIC} = 2k - 2(\ln)L \tag{5}$$

$$\text{BIC} = -2(\ln)L - K(\ln)n \tag{6}$$

K represents the number of parameters estimated in the model, $(\ln)L$ represents the maximized value of the log-likelihood function for the estimated model and n represents the number of observations in the dataset.

3. Data Analysis and Interpretation

3.1. Exploratory Data Analysis

The distribution of lung cancer status based on gender as display in **Figure 1**

indicates that there is a high prevalence of lung cancer among males compared to females.

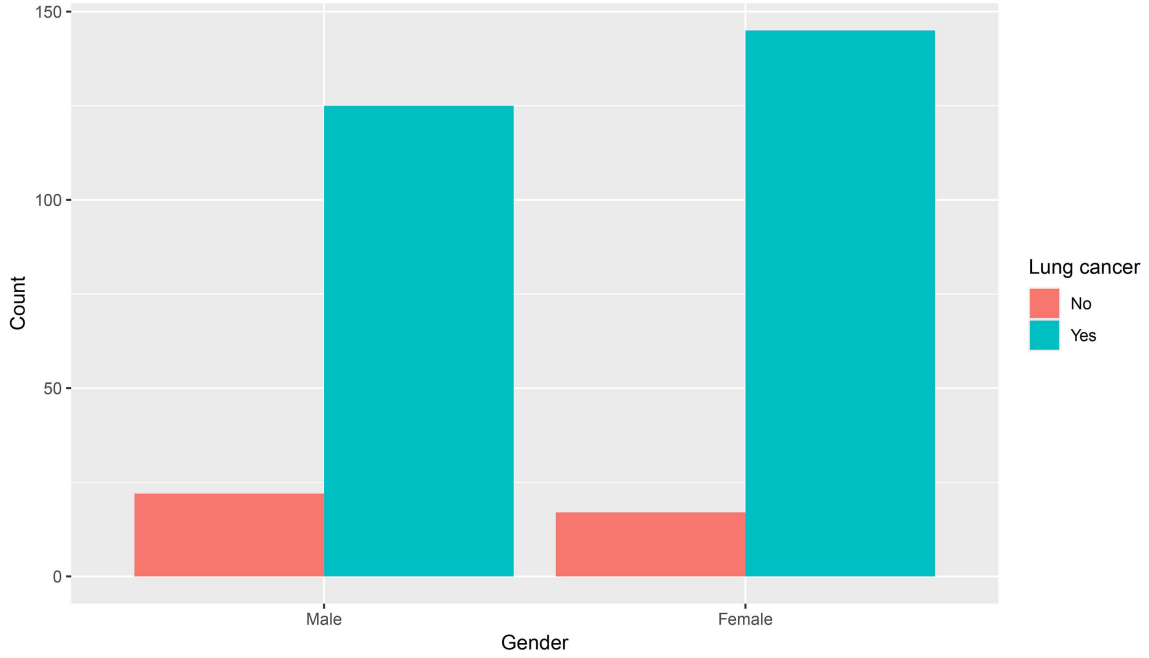

Figure 1. The relationship between gender and lung cancer.

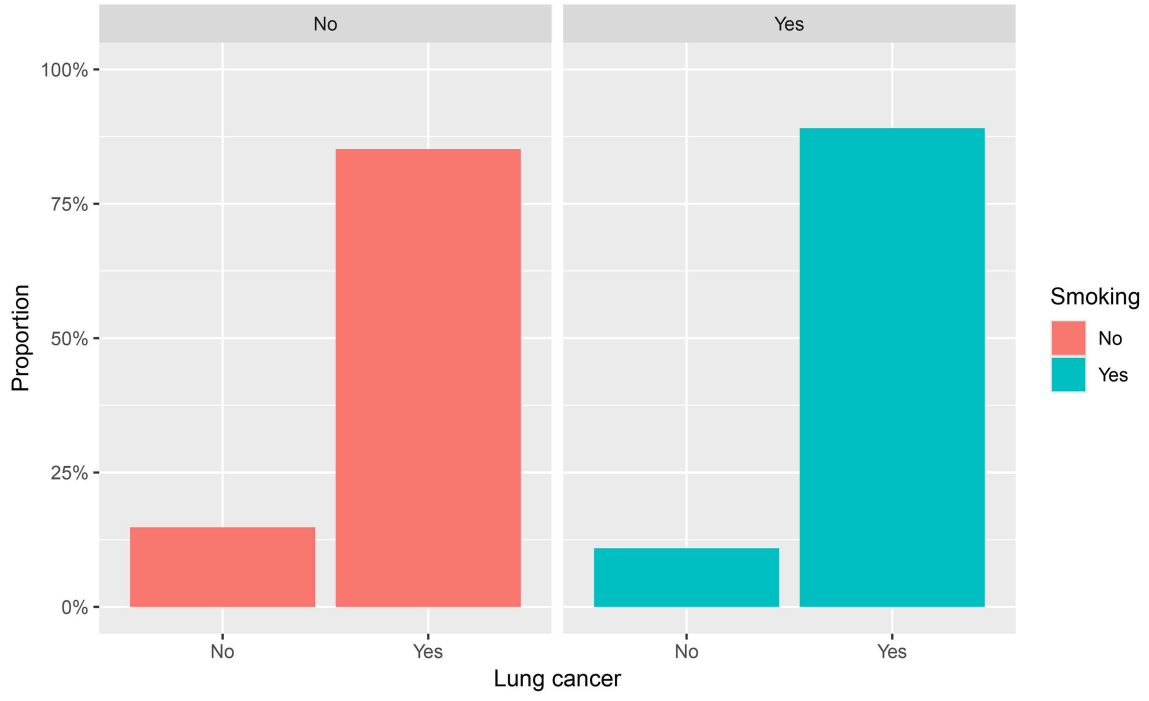

Figure 2. The relationship between smoking and lung cancer.

The proportion of patients who smoke with lung cancer is higher compared to those who do not smoke as shown in **Figure 2**.

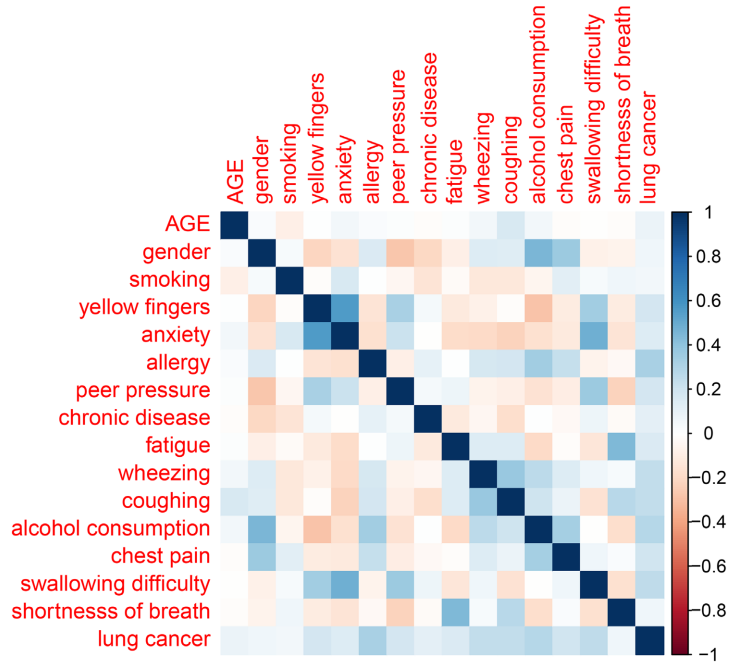

Figure 3. Correlation heatmap.

The correlation heatmap (**Figure 3**) shows the degree of relationship between variables. This revealed that there's a weak positive relationship between patients who have lung cancer and the lifestyle factors that may contribute to lung cancer.

3.2. Model Diagnostic

Figure 4 and **Figure 5** show the predicted probability plots of the logit and probit models, respectively. The plots indicate the model's good performance. However, there is barely a difference between the two plots, so the better-performing model cannot be easily detected.

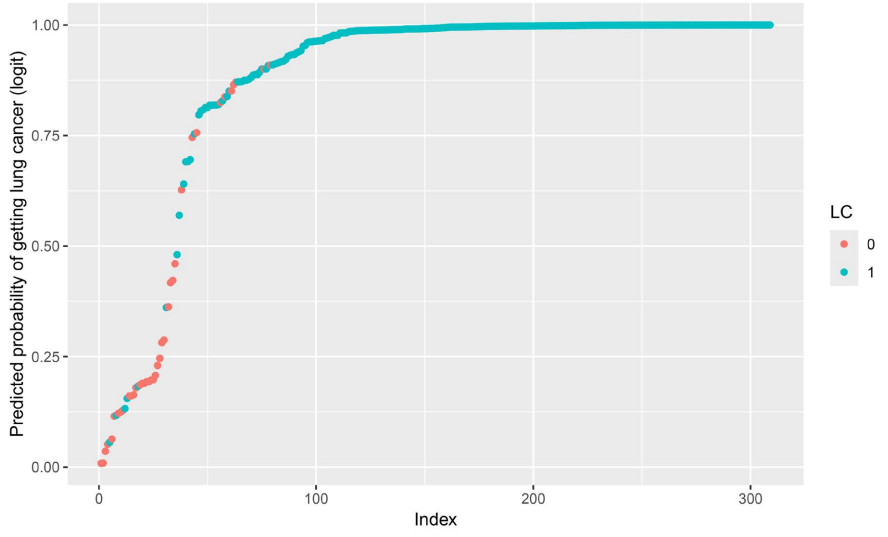

Figure 4. Predicted probability plot from logit model.

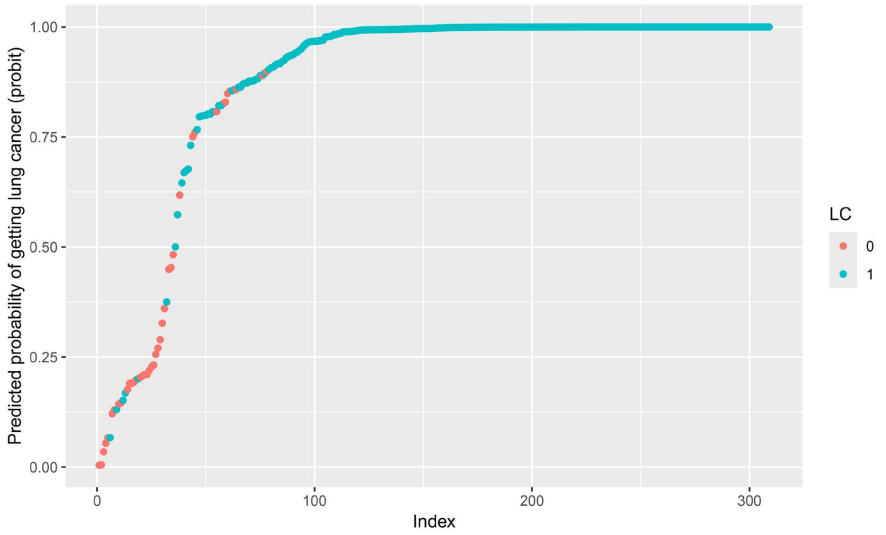

Figure 5. Predicted probability plot from probit model.

Table 1 displays the metric table, probit precision, recall, F1 score, indicating that the probit model is closer to 1 than the logit model. Also, the AIC and BIC of the probit is smaller which indicates that the probit model fits the data better.

Table 1. Model evaluation metric.

Metrics	Logit	Probit
AIC	123.91	122.65
BIC	183.643	182.379
Precision	0.9634	0.9635
Recall	0.9741	0.9778
F1 score	0.9687	0.9706

Table 2 shows the confusion matrix of the logit and probit models. When compared, the probit model has more true positives, indicating a better fit.

Table 2. Confusion matrix.

Model	Predicted	Actual	
		No	Yes
Logit	No	29	7
	Yes	10	263
Probit	No	29	6
	Yes	10	264

The receiver operating characteristics (ROC) curves for both logit and probit shown in Figure 6 and Figure 7 indicate a good performance of the model. However, the AUC of probit is closer to 1 which indicates it fits the model better.

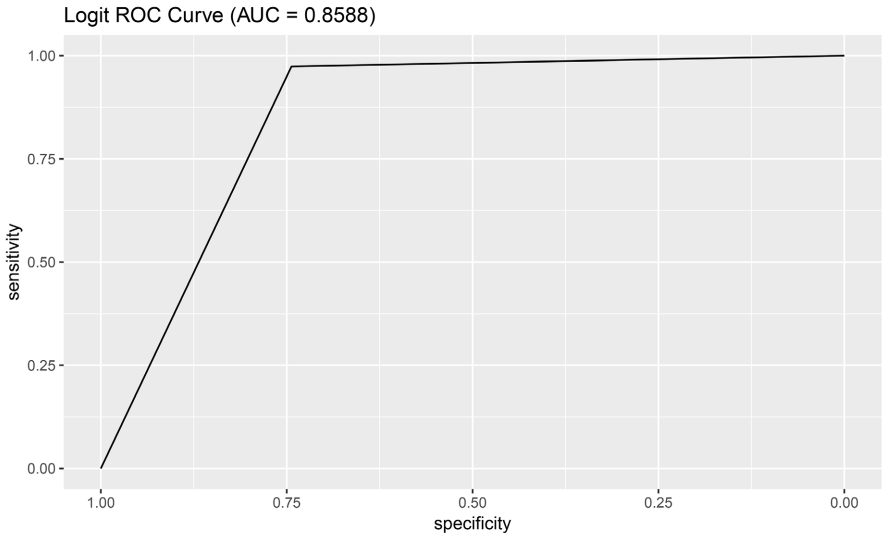

Figure 6. Logit ROC curve.

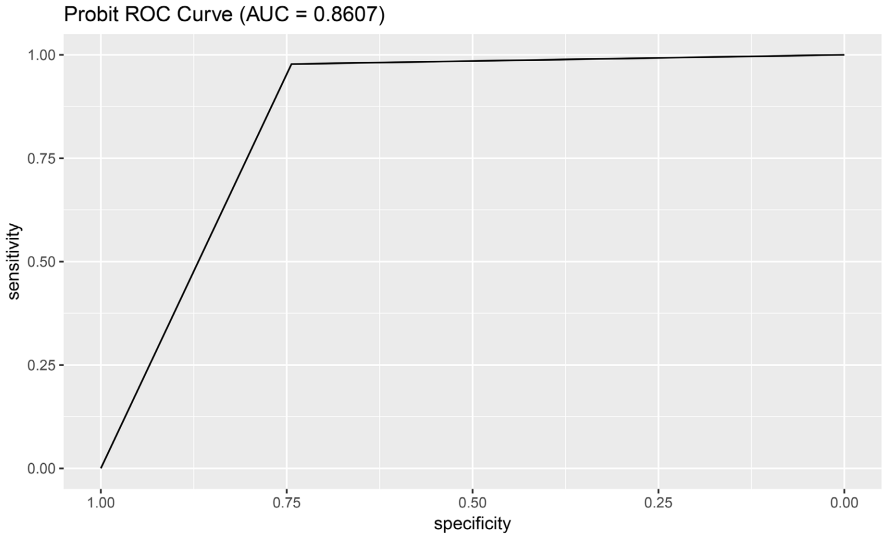

Figure 7. Probit ROC curve.

3.3. Model Summary

The analysis presents the results of classical logit and probit models alongside their corresponding Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC) in **Table 2**. Notably, both models yield negative intercepts as shown in **Table 3** and **Table 4**. The Logit intercept of −8.3376 suggests that when all risk factors and symptoms are absent (*i.e.* age is constant, the patient is not male, does not smoke, does not have yellow fingers, and other factors are absent) the odds of being diagnosed with lung cancer becomes extremely low, at approximately 0.024%. The Probit intercept of −4.5871 indicates a similar interpretation to that of the Logit intercept, when all risk factors and symptoms are absent, the probability of a patient developing the condition is extremely low, at approximately 1.02%.

Lifestyle factors like smoking emerge as significant determinants of lung cancer, underscoring their substantial impact. While alcohol consumption does not significantly influence the presence of lung cancer. Additionally, several health factors, including chronic diseases, fatigue, allergy, coughing, and swallowing difficulty, are identified as significant contributors to lung cancer occurrence.

Interestingly, while the probit model identifies the yellow finger as a significant factor, the logit model does not accord it the same significance.

When the results of the classical logit and probit models are compared, they closely align. However, when considering the specified factors, the AIC and BIC metrics favour the probit model as a superior predictor of lung cancer incidence.

Table 3. Summary of the logit models.

Factors				Confidence interval	
	Estimate	Standard error	p-value	2.5%	97.5%
Intercept	−8.3376	2.5069	0.0008	−13.4816	−3.4553
Age	0.0218	0.0339	0.5205	−0.0510	0.0870
Gender (Male)	−0.5261	0.7090	0.4581	1.9900	0.8347
Smoking	1.7760	0.7019	0.0114	0.4617	3.2488
Yellow fingers	1.3764	0.7425	0.0638	−0.0151	2.9515
Anxiety	0.8878	0.8127	0.2747	−0.6709	2.5553
Chronic disease	3.1916	0.8883	0.0003	1.5988	5.1279
Fatigue	3.0704	0.8252	0.0002	1.5927	4.8685
Allergy	1.6461	0.7689	0.0323	0.1850	3.2381
Wheezing	0.9663	0.8342	0.2467	−0.7218	2.6025
Alcohol	1.4098	0.7989	0.7762	−0.1254	3.0513
Coughing	3.3113	1.0717	0.0020	1.3404	5.5925
Shortness of breath	−0.7289	0.7600	0.3376	−2.2632	0.7608
Peer pressure	1.7312	0.6603	0.0087	0.5164	3.1430
Swallowing difficulty	3.1221	1.1298	0.0057	1.0960	5.5928
Chest pain	0.5591	0.6891	0.4172	−0.8227	1.9254

Table 4. Summary of the probit models.

Factors	Estimate	Standard error	p-value	Confidence interval	
				2.5%	97.5%
Intercept	−4.5871	1.3813	0.0009	−7.3021	−2.0765
Age	0.0105	0.0190	0.5818	−0.0251	0.0467
Gender (Male)	−0.2656	0.3837	0.4888	−1.0482	0.4904
Smoking	1.0223	0.3853	0.0080	0.2773	1.8220
Yellow fingers	0.8083	0.4051	0.0460	0.0297	1.6291
Anxiety	0.4860	0.4431	0.2728	−0.3641	1.3823

Continued

Chronic disease	1.7911	0.4789	0.0002	0.8870	2.8457
Fatigue	1.7557	0.4443	7.76e−5	0.9318	2.6965
Allergy	0.9002	0.4127	0.0292	0.0988	1.7555
Wheezing	0.5573	0.4471	0.2126	−0.3397	1.4459
Alcohol	0.8169	0.4352	0.0605	−0.0344	1.6993
Coughing	1.8185	0.5641	0.0013	0.7340	3.0213
Shortness of breath	−0.4249	0.4109	0.3011	−1.2522	0.3744
Peer pressure	1.0032	0.3621	0.0056	0.3245	1.7590
Swallowing difficulty	1.7200	0.6075	0.0046	0.5919	3.0489
Chest pain	0.2909	0.3662	0.4270	−0.4439	1.0112

4. Discussion

In line with existing research from earlier studies, smoking, chronic illness, exhaustion, alcohol use, coughing, and trouble swallowing are consistently identified by models as important predictors of lung cancer risk. Smoking, in particular, stands out as the primary risk factor for lung cancer, with a robust dose-response relationship between smoking intensity and lung cancer incidence [10] [11]. Similarly, the presence of chronic diseases such as chronic obstructive pulmonary disease (COPD) and other respiratory conditions has consistently been linked to an increased risk of lung cancer [12].

The significant association between fatigue and lung cancer risk observed in the study aligns with numerous studies reporting fatigue as a common symptom in lung cancer patients, often preceding diagnosis [13]. Moreover, the established link between alcohol consumption and lung cancer risk is supported by meta-analytical evidence indicating a positive dose-response relationship between alcohol intake and lung cancer risk [14].

The findings also show differences between the classical logit and classical probit models for determining which elements are important indicators of lung cancer risk. The classical logit model, for example, did not find allergies to be a relevant factor, whereas the classical probit model did. Such differences may stem from the distinct underlying assumptions and estimation techniques employed by these models [15].

In addition, the study's information criteria (AIC and BIC) indicate that the classical probit model performed better in prediction than the classical logit model. This observation aligns with previous research indicating that probit models may outperform logit models under certain circumstances, particularly when the assumption of normality in the underlying error distribution is more appropriate [16] [17].

The probit model yielded a higher AUC value in the study. The Area Under the ROC Curve (AUC) serves as a metric to assess the discriminatory ability of these models [18]. The probit model's higher AUC value of 0.8607 suggests that it exhibits a stronger discriminatory capability in predicting lung cancer outcomes

within our dataset compared to the logit model which had an AUC value of 0.8588.

The study outcome demonstrates that the classical probit model predicts lung cancer more accurately based on patient data and lifestyle factors than the classical logit model. More studies are necessary to confirm these results in wider populations and investigate other potential risk factors for lung cancer.

5. Conclusion

The study predicts the likelihood of having cancer-based on patients' data and lifestyle factors using probit and logit regression models. The significant factors from both models were smoking, chronic disease, fatigue, alcohol consumption, smoking difficulty, and coughing. The AIC of the probit model shows a superior predictive ability over the logit regression model. This implies that the probit model is more appropriate in predicting lung cancer risk. In addition, a great deal of research needs to be done to validate the model's predictive accuracy by expanding the categories or lifestyle factors of each variable in the datasets for the patients. Awareness campaigns should also be implemented and promoted for the target audience of lung cancer, with a particular focus on these factors—chronic diseases, alcohol, smoking, and respiratory symptoms—to encourage early detection of the disease.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Osmani, L., Askin, F., Gabrielson, E. and Li, Q.K. (2018) Current WHO Guidelines and the Critical Role of Immunohistochemical Markers in the Subclassification of Non-Small Cell Lung Carcinoma (NSCLC): Moving from Targeted Therapy to Immunotherapy. *Seminars in Cancer Biology*, **52**, 103-109. https://doi.org/10.1016/j.semcancer.2017.11.019

[2] Goldstraw, P., Chansky, K., Crowley, J., Rami-Porta, R., Asamura, H., Eberhardt, W.E.E., *et al.* (2016) The IASLC Lung Cancer Staging Project: Proposals for Revision of the TNM Stage Groupings in the Forthcoming (Eighth) Edition of the TNM Classification for Lung Cancer. *Journal of Thoracic Oncology*, **11**, 39-51. https://doi.org/10.1016/j.jtho.2015.09.009

[3] Alberg, A.J., Brock, M.V., Ford, J.G., Samet, J.M. and Spivack, S.D. (2013) Epidemiology of Lung Cancer. *Chest*, **143**, e1S-e29S. https://doi.org/10.1378/chest.12-2345

[4] Hamra, G.B., Guha, N., Cohen, A., Laden, F., Raaschou-Nielsen, O., Samet, J.M., *et al.* (2014) Outdoor Particulate Matter Exposure and Lung Cancer: A Systematic Review and Meta-Analysis. *Environmental Health Perspectives*, **122**, 906-911. https://doi.org/10.1289/ehp.1408092

[5] Hosmer, D.W., Lemeshow, S. and Sturdivant, R.X. (2013) Applied Logistic Regression. Wiley. https://doi.org/10.1002/9781118548387

[6] Peng, C.J. and So, T.H. (2002) Logistic Regression Analysis and Reporting: A Primer. *Understanding Statistics*, **1**, 31-70. https://doi.org/10.1207/s15328031us0101_04

[7] Biswas, A. and Nath. A. (2024) Lung Cancer Dataset. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/8795028

[8] Karabiber, F. (2024) Binary Variable—LearnDataSci. https://www.learndatasci.com/glossary/binary-variable

[9] Hosmer Jr., D.W., Lemeshow, S. and Sturdivant, R.X. (2013) Applied Logistic Regression. John Wiley and Sons.

[10] Pesch, B., Kendzia, B., Gustavsson, P., Jöckel, K., Johnen, G., Pohlabeln, H., *et al.* (2011) Cigarette Smoking and Lung Cancer—Relative Risk Estimates for the Major Histological Types from a Pooled Analysis of Case-Control Studies. *International Journal of Cancer*, **131**, 1210-1219. https://doi.org/10.1002/ijc.27339

[11] Thun, M.J., Carter, B.D., Feskanich, D., Freedman, N.D., Prentice, R., Lopez, A.D., *et al.* (2013) 50-Year Trends in Smoking-Related Mortality in the United States. *New England Journal of Medicine*, **368**, 351-364. https://doi.org/10.1056/nejmsa1211127

[12] Durham, A.L. and Adcock, I.M. (2015) The Relationship between COPD and Lung Cancer. *Lung Cancer*, **90**, 121-127. https://doi.org/10.1016/j.lungcan.2015.08.017

[13] Bower, J.E., Ganz, P.A., Desmond, K.A., Bernaards, C., Rowland, J.H., Meyerowitz, B.E., *et al.* (2006) Fatigue in Long-Term Breast Carcinoma Survivors. *Cancer*, **106**, 751-758. https://doi.org/10.1002/cncr.21671

[14] Bagnardi, V., Rota, M., Botteri, E., Tramacere, I., Islami, F., Fedirko, V., *et al.* (2014) Alcohol Consumption and Site-Specific Cancer Risk: A Comprehensive Dose-Response Meta-Analysis. *British Journal of Cancer*, **112**, 580-593. https://doi.org/10.1038/bjc.2014.579

[15] Winkelmann, R. and Boes, S. (2006) Analysis of Microdata. Springer.

[16] Cramer, J.S. (2003) Logit Models from Economics and Other Fields. Cambridge University Press. https://doi.org/10.1017/cbo9780511615412

[17] Maddala, G.S. (1983) Limited-Dependent and Qualitative Variables in Econometrics. Cambridge University Press. https://doi.org/10.1017/cbo9780511810176

[18] Jiménez-Valverde, A. (2011) Insights into the Area under the Receiver Operating Characteristic Curve (AUC) as a Discrimination Measure in Species Distribution Modelling. *Global Ecology and Biogeography*, **21**, 498-507. https://doi.org/10.1111/j.1466-8238.2011.00683.x