LUNG CANCER DISEASE PREDICTION MODEL WITH MULTIPLE LINEAR REGRESSION

Ratna Puspita Indah[1*], Anisatul Farida[2], Sharifah Noor Masidayu[3]
Duta Bangsa University[1,2], Multimedia University Malaysia[3]
*Correspondence Email : ratna_puspita@udb.ac.id

ABSTRACT

Lung cancer remains one of the leading causes of cancer-related mortality worldwide, particularly in developing countries where smoking prevalence is high. This study aims to develop a predictive model for lung cancer risk using multiple linear regression based on two main factors: genetic predisposition and exposure to passive smoking. The research was conducted using an observational analytic design with secondary data derived from cancer registries, hospital medical records, and national health surveys. Data processing included cleaning, imputation of missing values, and standardization of exposure variables. The results of the regression analysis showed that both genetic risk and passive smoking significantly increased the lung cancer risk score, with coefficients of 0.24 and 0.48, respectively. Interestingly, passive smoking demonstrated a stronger impact compared to genetic predisposition, indicating its role as a more dominant determinant of lung cancer risk. The model explained 20.5% of the variation in risk, while the remaining was influenced by other factors such as air pollution, occupational exposure, and lifestyle. These findings highlight the importance of strengthening public health policies, particularly tobacco control in public spaces, and implementing targeted risk-based screening strategies. This predictive model offers a practical tool for early detection, efficient allocation of health resources, and effective cancer prevention strategies.

KEYWORDS Lung cancer, multiple linear regression, genetic risk, passive smoking, risk prediction.

INTRODUCTION

Lung cancer remains a major global health problem. In 2022, approximately 2.5 million new cases were recorded. More than 1.8 million people died from lung cancer that same year. Lung cancer is the most common cancer worldwide (Arsunan, 2023).

Furthermore, this disease is also the leading cause of cancer deaths globally. This situation highlights the need for better strategies, including prevention, early detection, and risk prediction. In Indonesia, the situation is no less serious. Lung cancer ranks second in terms of cancer incidence. However, it is the leading cause of cancer deaths. The high mortality rate indicates that many cases are detected at an advanced stage. Early symptoms of lung cancer are often nonspecific, and awareness of screening remains low (Dianita et al., 2025).

The primary cause of lung cancer is tobacco. Approximately 85 percent of cases are linked to tobacco use. Environmental cigarette smoke increases the risk in passive smokers. Besides tobacco, air pollution is also a significant factor. Fine particulate matter, such as PM2.5, is associated with lung cancer. The risk even increases in nonsmokers. Exposure to PM2.5 can trigger tumor development in lung tissue(Sinaga et al., 2024). Recent studies have shown an interaction with latent mutations. Inflammatory processes also play a role in carcinogenesis. Occupational factors also increase the risk. Exposure to asbestos, diesel exhaust, and welding fumes are examples of occupational risks. Radon is also a known cause of lung cancer. Household biomass burning increases the risk. Poor home ventilation exacerbates the situation(Basuki et al., 2025).

Inequality in healthcare access worsens patient outcomes. Many people receive diagnosis and treatment late. This reduces patients' chances of survival. The burden on families and the healthcare system increases. Risk prediction strategies are crucial. Risk prediction helps identify high-risk individuals (Edoh et al., 2024). This supports more targeted screening programs. Limited healthcare resources can be allocated more efficiently.

In epidemiological research, multiple linear regression is widely used. This method relates a continuous outcome variable to a number of predictors. The outcome can be a risk score or the probability of lung cancer (Dritsas & Trigka, 2022). The advantage of linear regression is its clear interpretation. The resulting coefficients are easy for clinicians and policymakers to understand. Simple interpretation facilitates risk communication. However, there are challenges with multicollinearity between variables, for example, between smoking intensity, smoking duration, and cumulative exposure. Therefore, variable selection is crucial.

Data quality significantly impacts model results. Ideally, data should include demographics, smoking history, and air pollution. Employment history should also be recorded. Comorbidities such as chronic lung disease influence risk. Environmental indicators such as PM2.5 concentration can be added (Wang & Liu, 2023). Geospatial information strengthens model accuracy. Model validation is necessary for reliable results. External validation ensures the model's applicability across different regions. Model performance is assessed through calibration and discrimination. Good calibration prevents both overestimation and underestimation of risk.

Multiple linear regression is also easy to implement. Simple models are suitable for developing countries like Indonesia. They require low computational requirements, making them efficient. Transparent coefficients facilitate policy evaluation. The burden of lung cancer increases with population aging. Urbanization worsens air quality. Southeast Asia faces a combination of high risks. Predictive models must be tailored to the local context. Exposure to biomass smoke remains prevalent in rural areas. The interaction of risk factors needs to be analyzed(Mitra et al., 2025).

Linear regression allows testing for variable interactions. Nonlinear relationships can be addressed with simple transformations (James et al., 2023). Outcome definitions must be clear and consistent. Measurement of exposure variables must be precise. Tobacco control policies remain a priority. Air pollution reduction brings significant benefits. Risk prediction supports population health policies. Multiple linear regression forms the basis

of initial modeling. Simple models remain important despite the availability of sophisticated methods. Overall, transparent and contextualized predictive models are essential (Busari & Bolanle, 2025).

RESEARCH METHOD

1. Type and Design of Research

This study is an observational study with an analytical design. These studies are able to show the incidence of a disease or outcome, and the relationship between exposure and outcome is usually expressed as relative risk. They are valuable for assessing causality, though high dropout rates among participants and the presence of confounding factors may pose challenges (Ramji, 2022), where data were obtained from secondary sources such as cancer registries, hospital medical records, and national health surveys. This design was chosen to explore the relationship between predictor variables (risk factors) and the dependent variable (lung cancer risk).

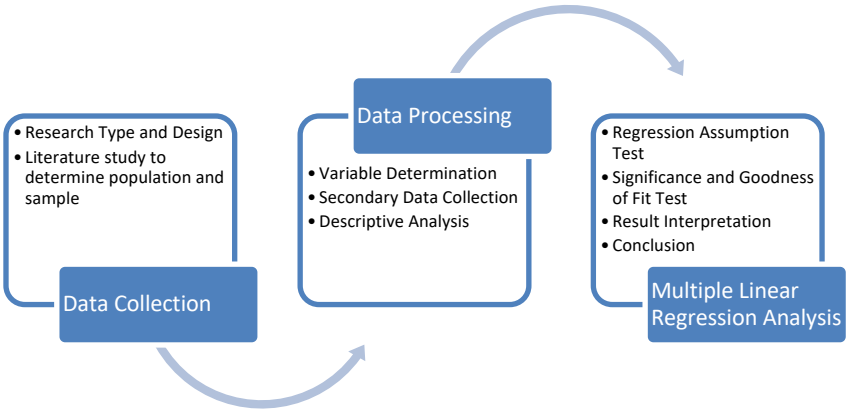

Figure 1. Research Method

2. Population and Sample

The study population included individuals with documented health data related to lung cancer risk factors. The sampling technique was purposive, with inclusion criteria being the availability of complete data, including smoking history, pollution exposure, and demographic variables. Data with incomplete information, inconsistencies, or extreme outliers were excluded from the analysis(Berger & Kiefer, 2021).

3. Research Variable

a. **Dependent Variable (Y):** lung cancer risk score in continuous form (e.g., probability of occurrence score).
b. **Independent Variable (X):** genetic risk is the possibility that a person will develop certain conditions or diseases based on their genetic makeup and passive smokers are passive smokers.

4. Data Collection and Processing

The obtained data will undergo a data cleaning process to eliminate duplication and correct input errors (Ilyas & Rekatsinas, 2022). Missing values will be handled using appropriate imputation techniques (Farida et al., 2025). Next, variable transformations will be performed, such as calculating the cumulative smoking exposure index (pack-years) or standardizing air pollution exposure by location.

5. Descriptive Analysis

Descriptive analysis was conducted to describe the sample characteristics, both demographic and exposure to risk factors. The results of the analysis are presented in the form of distribution tables, diagrams, and measures of central tendency. Accordingly, external factors include government initiatives, technology, and social media. To address this, marketing strategies have been redesigned to reduce potential risks faced by consumers. Therefore, as digital and technological trends define consumer behavior, businesses must integrate digital transformation within their practices (Cruz-Cárdenas et al., 2021).

6. Linear Regression Assumtion Test

Prior to the main analysis, regression assumption tests were conducted, including:

a. Normality of residuals to ensure the error distribution follows a normal distribution.
b. Multicollinearity to detect high correlations between predictor variables.
c. Homoscedasticity to check for equality of error variances.
d. Identification of outliers and leverage points that may affect model stability.

7. Multiple Linear Regression Analysis

The prediction model is built with multiple linear regression using the formula:

$$\hat{Y} = a_0 + a_1\hat{X}_1 + a_2\hat{X}_2$$

with :

$$\sum x_1 y = a_1 \sum x_1^2 + a_2 \sum x_1\, x_2$$
$$\sum x_2\, y = a_1 \sum x_1\, x_2 + a_2 \sum x_2^2$$

To determine value a_0, a_1 dan a_2

Regression coefficient (α) shows the magnitude of the influence of each risk factor on the lung cancer risk score.

8. Significance test and Goodness of Fit

The *t-test* is used to test the partial influence of each predictor variable.

$$t_{count} = \frac{r_{y2.1}\sqrt{n-3}}{\sqrt{(1-r_{y2.1}^2)}}, \quad t_{table}(\alpha; n-k-1)$$

The *F test* is used to assess the overall significance of the model.

$$F_{count} = \frac{R_{y.12}^2(n-k-1)}{k(1-R^2)}, F_{tabel}(\alpha; \frac{db(Reg)}{db(S)})$$

The *R² value* is used to measure the proportion of variation in lung cancer risk that can be explained by the model.

$$r_{y1.2} = \frac{r_{y1} - r_{y2} \times r_{12}}{\sqrt{(1-r_{y2}^2) - (1-r_{12}^2)}}$$

9. Interpretation and Discussion

The analysis results are interpreted to determine the dominant risk factors for lung cancer. Discussions are conducted comparing the results with previous studies and outlining the implications of the findings for prevention strategies and health policy.

10. Conclusion

The study's conclusion confirms the multiple linear regression model obtained, the significant variables influencing lung cancer risk, and its potential application in risk-based screening and intervention programs.

RESULT AND DISCUSSION

The results of the multiple linear regression analysis indicate that the Genetic Risk (X_1) and Passive Smoker (X_2) factors have a positive influence on the lung cancer risk score. A constant value of 0.39 indicates that even though both predictor variables are at their lowest levels, individuals still have a baseline risk of lung cancer. This is in line with the concept that other risk factors outside the model, such as air pollution or occupational exposure, still play a role in triggering lung cancer.

$$\sum x_1 y = a_1 \sum x_1^2 + a_2 \sum x_1 x_2$$
$$\sum x_2 y = a_1 \sum x_1 x_2 + a_2 \sum x_2^2$$
$$46 = 226a_1 + 167a_2$$
$$24 = 167a_1 + 330a_2$$
$$\begin{bmatrix} 226 & 167 \\ 167 & 330 \end{bmatrix} a_1 = \begin{bmatrix} 46 & 167 \\ 24 & 330 \end{bmatrix}$$
$$46.691a_1 = 11.172 \rightarrow a_1 = 0{,}24$$
$$\begin{bmatrix} 226 & 167 \\ 167 & 330 \end{bmatrix} a_2 = \begin{bmatrix} 226 & 46 \\ 167 & 24 \end{bmatrix}$$
$$46.691a_2 = 22.258 \rightarrow a_2 = 0{,}48$$
$$a_0 = \bar{y} - a_1\overline{x_1} - a_2\overline{x_2}$$
$$a_0 = 3{,}2 - (0{,}24 \times 5{,}27) - (0{,}48 \times 4{,}85) = 0{,}39$$

Thus, the regression equation model of Y on X_1 *and* X_2 can be formulated as follows:

$$\bar{Y} = 0{,}39 + 0{,}24x_1 + 0{,}48x_2$$

Interpretation :

a. If Genetic Risk (X1) and Passive Smoker (X2) are close to or equal to 0, then the cancer risk score (Y) = 0.39. With the interpretation that if Genetic Risk (X1) and Passive Smoker (X2) are low, then the average Cancer Risk Score (Y) remains at 0.39. This means that the overall Cancer Risk Score (Y) remains the same.
b. If Genetic Risk (X1) increases by one unit, while Passive Smoker (X2) remains the same, then the cancer risk score (Y) = 0.24 times. With the interpretation that every increase in Genetic Risk (X1) by one unit will have an impact on increasing the Cancer Risk Score (Y) by 0.24.
c. If Passive Smoker (X2) increases by one unit, while Genetic Risk (X1) remains the same, then the Cancer Risk Score (Y) will increase by 0.48 times. With the interpretation that each increase in Passive Smoker (X2) will have an impact on increasing the Cancer Risk Score (Y) by 0.48.

Specifically, a one-unit increase in Genetic Risk (X_1) will increase the risk score by 0.24, while a one-unit increase in Passive Smoker (X_2) will increase the risk score by

0.48. These results imply that the role of secondhand smoke exposure is more dominant than genetic predisposition in the context of this study. This finding is consistent with previous studies that stated that environmental cigarette smoke exposure contributes to more than 20% of lung cancer cases in non-smokers, especially in developing countries with high smoking prevalence (Cheng et al., 2021). This suggests that many non-smokers are diagnosed at advanced stages of the disease, when treatment becomes more difficult. The study highlights the importance of advancing early detection approaches, informed by research from 2020 to 2024 as well as practical insights from the Republican Specialized Scientific and Practical Medical Centre of Oncology and Radiology in Tashkent, Uzbekistan. Effective lung cancer prevention requires addressing the full spectrum of risk factors, not merely those linked to smoking (Alimkhodjayeva et al., 2025).

Biologically, exposure to secondhand smoke contains thousands of harmful chemicals, including carcinogens such as benzene, arsenic, and nitrosamines, which can trigger DNA damage and chronic inflammation in lung tissue (Dehghani et al., 2024). This explains why the regression coefficient for the Passive Smoker variable is greater than that for Genetic Risk. Meanwhile, genetic factors play a predisposing role through certain germline mutations, for example in the EGFR or TP53 genes, which increase an individual's susceptibility to environmental carcinogens.

This increasing risk trend is also consistent with the multiple-hit carcinogenesis theory, which states that cancer generally results from the accumulation of genetic damage exacerbated by environmental factors (Weeden et al., 2023). Thus, the combination of genetic predisposition and exposure to secondhand smoke increases the likelihood of lung cancer.

$$F_{hitung} = \frac{RJK(Reg)}{RJKS} = \frac{11{,}28}{0{,}64} = 17{,}625$$

$$F_{tabel}\left(\alpha; \frac{db(Reg)}{db(S)}\right) \rightarrow F_{tabel}\left(0{,}05; \frac{2}{57}\right) = 3{,}16$$

The results of the Deviation from Linearity analysis obtained an F of 17.625 with an F table value of 3.16. Because $F_{count} > F_{table}$ then reject Ho and accept H1. Thus, the correlation between x and y is positive and significant. The Deviation from Linearity test produced an F-value of 17.625, while the critical value from the F-table was 3.16. Since the calculated F (17.625) is greater than the table value (3.16), the null hypothesis (Ho), which assumes no significant relationship, is rejected. Instead, the alternative hypothesis (H1) is accepted. This indicates that there is a statistically significant and positive correlation between variable X and variable Y (Essam et al., 2022).

The results of the Deviation from Linearity analysis obtained an F of 17.625 with an F table value of 3.16. Because $F_{count} > F_{table}$ then reject Ho and accept H1. Thus, the correlation between x and y is positive and significant. The Deviation from Linearity test resulted in an F-value of 17.625, compared to the F-table value of 3.16. Since the calculated F (F_count) is greater than the critical value (F_table), the null hypothesis (Ho) is rejected and the alternative hypothesis (H1) is accepted. This means that the relationship between X and Y is not only positive but also statistically significant, showing that changes in X are meaningfully associated with changes in Y

Coefficient of Determination

$$R^2_{y.12} = \frac{JK(Reg)}{\sum y^2} = \frac{22.56}{110} = 0{,}205$$
$$KD = R^2_{y.12} \times 100\% = 0{,}205 \times 100\% = 20{,}5\%$$

The coefficient of determination obtained R square 0.25. Thus, the coefficient of determination of KD obtained 20.5%. So, 20.5% of the variation in the dependent variable can be explained by variables x_1 and x_2 and the rest by other factors not studied. However, based on the results reported, the KD value is 20.5%. This means that approximately 20.5% of the total variation in the dependent variable (Y) can be explained by the independent variables x_1 and x_2 The remaining 79.5% of the variation is due to other variables or factors not included in the model

Compared with previous research, these results reinforce the report by (Hecht et al., 2021) which emphasized that individuals with a family history of lung cancer have a relatively higher risk, but tobacco smoke exposure remains the strongest determinant. Similarly, the World Health Organization (2022) reported that exposure to secondhand smoke increases the risk of lung cancer by 20–30%, particularly in Southeast Asian populations with a high prevalence of active smoking.

These findings have important implications for public health strategies. First, intervention programs need to strengthen smoking control policies in public spaces to reduce secondhand smoke exposure. Second, risk-based screening can be targeted to individuals with a genetic history of lung cancer for earlier detection. Third, the results of this predictive model can form the basis for developing an early warning system that integrates environmental, genetic, and lifestyle factors.

Overall, the multiple linear regression model in this study proved effective in explaining variations in lung cancer risk based on genetic factors and secondhand smoke exposure. These results support the study's hypothesis that both variables significantly influence lung cancer risk scores, while also underscoring the urgency of stricter tobacco control policies in Indonesia.

CONCLUSION

This study demonstrates that genetic risk and passive smoking significantly influence the increase in lung cancer risk scores. The multiple linear regression model produced a constant of 0.39, with regression coefficients of 0.24 for genetic factors and 0.48 for passive cigarette smoke exposure. This confirms that passive cigarette smoke exposure has a more dominant influence than genetic predisposition in increasing lung cancer risk.

The coefficient of determination indicates that the model can explain 20.5% of the variation in risk, with the remainder influenced by other unexplored factors, such as air pollution, the work environment, and lifestyle. These findings align with previous studies that emphasized the dangers of secondhand smoke exposure as a major determinant of lung cancer in developing countries with a high prevalence of active smoking.

Practically, the results of this study emphasize the importance of tobacco control policies in public spaces, the implementation of risk-based screening for individuals with a genetic history, and the development of an early warning system based on environmental and lifestyle factors. Therefore, this predictive model can form the basis for more effective public health interventions to reduce morbidity and mortality from lung cancer.

REFERENCES

Alimkhodjayeva, L. T., Norbekova, M. H., Kurbankulov, U. M., Khusainova, M. J., & Otajonov, J. H. (2025). *LUNG CANCER IN NON-SMOKERS: EMERGING RISK FACTORS AND CHALLENGES*.

Arsunan, A. (2023). *Karakteristik dan Luaran Penderita Kanker yang Terpapar Covid-19 di Rumah Sakit Pendidikan Unhas Tahun 2020-2022= Characteristics and Outcomes On Cancer Patients Exposed to Covid-19 In Unhas Hospital, 2020-2022*. Universitas Hasanuddin.

Basuki, R., Musyahidah, M., Risnawati, A., & Sumiati, B. (2025). *Kesehatan Lingkungan dan Kesehatan Kerja: Paparan, Risiko, dan Strategi Mitigasi*. PT Mafy Media Literasi Indonesia.

Berger, A., & Kiefer, M. (2021). Comparison of different response time outlier exclusion methods: A simulation study. *Frontiers in Psychology*, *12*, 675558.

Busari, M., & Bolanle, T. (2025). *INTEGRATING SOCIAL DETERMINANTS INTO PREDICTIVE MODELS FOR US PUBLIC HEALTH FORECASTING*.

Cheng, E. S., Weber, M., Steinberg, J., & Yu, X. Q. (2021). Lung cancer risk in never-smokers: An overview of environmental and genetic factors. *Chinese Journal of Cancer Research*, *33*(5), 548.

Cruz-Cárdenas, J., Zabelina, E., Guadalupe-Lanas, J., Palacio-Fierro, A., & Ramos-Galarza, C. (2021). COVID-19, consumer behavior, technology, and society: A literature review and bibliometric analysis. *Technological Forecasting and Social Change*, *173*, 121179.

Dehghani, M. H., Bashardoust, P., Nayeri, D., Ghalhari, M. R., Yazdi, N. B., Jajarmi, F., Karri, R. R., & Mubarak, N. M. (2024). A comprehensive review of the potential outcomes of exposure to tobacco smoke or secondhand smoke. *Health Effects of Indoor Air Pollution*, 167–189.

Dianita, E. M., Fuadiati, L. L., & Alizain, A. A. (2025). PENINGKATAN PENGETAHUAN TERHADAP DETEKSI DINI KANKER PAYUDARA PADA REMAJA. *Penerbit Tahta Media*.

Dritsas, E., & Trigka, M. (2022). Lung cancer risk prediction with machine learning models. *Big Data and Cognitive Computing*, *6*(4), 139.

Edoh, N. L., Chigboh, V. M., Zouo, S. J. C., & Olamijuwon, J. (2024). Improving healthcare decision-making with predictive analytics: A conceptual approach to patient risk assessment and care optimization. *International Journal of Scholarly Research in Medicine and Dentistry*, *3*(2), 1–10.

Essam, F., El, H., & Ali, S. R. H. (2022). A comparison of the pearson, spearman rank and kendall tau correlation coefficients using quantitative variables. *Asian J. Probab. Stat*, *20*(3), 36–48.

Farida, A., Atina, V., & Suwandi, D. (2025). Mathematical Modeling and Integration of Machine Learning-Based Prediction System on E-Learning Platform to Improve Students' Academic Performance. *JTAM (Jurnal Teori Dan Aplikasi Matematika)*, *9*(3), 829–839.

Hecht, C. A., Yeager, D. S., Dweck, C. S., & Murphy, M. C. (2021). Beliefs, affordances, and adolescent development: Lessons from a decade of growth mindset interventions. In *Advances in child development and behavior* (Vol. 61, pp. 169–197). Elsevier.

Ilyas, I. F., & Rekatsinas, T. (2022). Machine learning and data cleaning: Which serves the other? *ACM Journal of Data and Information Quality (JDIQ)*, *14*(3), 1–11.

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). Linear regression. In *An introduction to statistical learning: With applications in python* (pp. 69–134). Springer.

Mitra, P., Chakraborty, D., Nayek, S., Dan, U., & Mondal, N. K. (2025). The assessment of health risk among biomass smoke exposed rural tribal women and its effect on blood platelet activities. *Air Quality, Atmosphere & Health*, 1–14.

Ramji, S. (2022). Study design: observational studies. *Indian Pediatrics*, *59*(6), 493–498.

Sinaga, V. S. E. Z., Siahaan, P. G., Purba, N. R., Adilla, A., Harahap, A. P., Salya, V., Sitompul, J. G., & Tarigan, P. S. B. (2024). Analisis Persepsi Mahasiswa Terhadap Kawasan Tanpa Rokok Sebagai Manifestasi Hak Atas Kesehatan (Studi Kasus: Mahasiswa Pendidikan Biologi Kelas A). *Innovative: Journal Of Social Science Research*, *4*(6), 6654–6668.

Wang, Q., & Liu, S. (2023). The effects and pathogenesis of PM2. 5 and its components on chronic obstructive pulmonary disease. *International Journal of Chronic Obstructive Pulmonary Disease*, 493–506.

Weeden, C. E., Hill, W., Lim, E. L., Grönroos, E., & Swanton, C. (2023). Impact of risk factors on early cancer evolution. *Cell*, *186*(8), 1541–1563.