COMPUTATIONAL MATHEMATICAL MODELING FOR LUNG CANCER DISEASE PREDICTION USING MULTIPLE LINEAR REGRESSION

Anisatul Farida[1*], Ratna Puspita Indah[2], Dwi Hartanti[3] , Adão Manuel da Silva [4]
Universitas Duta Bangsa Surakarta[1,2,3], Instituto Superior Cristal, Timor Leste[4]
*Correspondence Email : anisatul_farida@udb.ac.id

ABSTRACT

Lung cancer remains one of the most prevalent and deadly types of cancer worldwide, especially in developing countries with high smoking rates and limited early detection resources. This study aims to develop a computational mathematical model for predicting lung cancer risk using multiple linear regression. The model focuses on two primary factors: genetic predisposition and exposure to passive smoking, which are among the most significant determinants of lung cancer. An observational analytic design was employed using secondary data obtained from cancer registries, hospital records, and national health survey datasets. Computational data preprocessing techniques, including data cleaning, missing value imputation, and variable normalization, were applied to ensure model accuracy and reliability. The regression analysis revealed that both genetic predisposition and passive smoking significantly increased the lung cancer risk score, with regression coefficients of 0.24 and 0.48, respectively. The findings indicate that passive smoking has a greater impact on lung cancer risk compared to genetic factors. The final model demonstrated a coefficient of determination (R^2) 0.72 indicates that 72% of the variation in risk can be explained by the combination of these two variables. This finding suggests that environmental factors have a more dominant influence than lifestyle factors on increasing lung cancer risk. This computational model provides a practical tool for early detection and risk stratification, supporting public health policies aimed at tobacco control and targeted screening programs to reduce lung cancer incidence and mortality.

KEYWORDS

Lung cancer, cLung cancer, multiple linear regression, passive smoking, genetic predisposition

INTRODUCTION

Lung cancer is one of the leading causes of cancer death worldwide and poses a serious challenge to the global health system. According to the World Health Organization (WHO) and the International Agency for Research on Cancer (IARC), by 2022, there will be an estimated 2.48 million new cases of lung cancer, resulting in 1.8 million deaths annually (Zhou et al., 2024). This figure makes lung cancer the deadliest type of cancer, accounting for approximately 18–19% of all cancer deaths worldwide. The burden of this disease tends to be higher in developing countries with high smoking prevalence, increasing levels of air pollution, and limitations in early detection facilities and health services. This condition results in many patients being diagnosed at an advanced stage, resulting in low cure rates and high mortality rates. The WHO also estimates that without appropriate intervention, the incidence of lung cancer will continue to rise in the coming decades, especially in resource-limited regions (Jiwnani et al., 2022).

This demonstrates the importance of prevention efforts, tobacco control, and the application of predictive technology to support early detection and more effective health policy planning. Early detection of lung cancer is often difficult due to nonspecific initial symptoms, resulting in most patients being diagnosed at an advanced stage(Kalinke et al., 2021). This condition results in low life expectancy and increases the economic burden on families and the nation. Therefore, an effective risk prediction approach is needed to support early detection and decision-making in health planning.

The development of computational mathematics and data analytics technology has provided new opportunities in the healthcare sector, particularly in data-driven predictive modeling. Computational mathematics combines mathematical theory, algorithms, and computation to solve complex problems that are difficult to solve analytically (Mitra et al., 2025). In the context of this research, computational mathematics is used to build a lung cancer risk prediction model capable of processing large amounts of data quickly and accurately. In this study, computational mathematics is used to build a lung cancer risk prediction model through several stages.

The process begins with data preprocessing, including cleaning, filling in missing values, and standardization to prepare the data for analysis (El Morr et al., 2022). Next, multiple linear regression (MLR) was used, calculated using the ordinary least squares (OLS) method, to determine the influence of genetic factors and cigarette smoke exposure on lung cancer risk. To ensure accurate results, regression assumptions were tested, including residual normality, multicollinearity, and heteroscedasticity. If problems such as overfitting were found, regularization methods such as Ridge or Lasso were applied. The model's reliability is then tested using cross-validation and bootstrap techniques, ensuring more stable prediction results. With the support of software such as Python or R, this method allows for rapid and efficient analysis of large amounts of data, producing predictive models that can be used for early detection and health policy planning. One method frequently used in predictive modeling is multiple linear regression (MLR), which allows for the analysis of the relationship between multiple independent variables and a single dependent variable (Etemadi & Khashei, 2021). MLR is effective in measuring the contribution of certain risk factors, such as genetic predisposition and secondhand smoke exposure, which are widely recognized as dominant factors in the development of lung cancer.

The selection of multiple linear regression (MLR) in this study was based on scientific and technical considerations relevant to the research objective, which is to predict lung cancer risk based on several key risk factors. Lung cancer is a disease influenced by various interacting factors, such as genetic predisposition, exposure to secondhand smoke,

air pollution, lifestyle, and other environmental factors (Cheng et al., 2021). The multiple linear regression method allows for the simultaneous analysis of the relationship between two or more independent variables to determine the extent to which each factor influences the dependent variable, namely the lung cancer risk score (Weisburd et al., 2021). In the context of this study, MLR is used to assign weights in the form of regression coefficients to each risk factor so that it can be identified which factor has the most dominant influence on the occurrence of lung cancer.

Another advantage of multiple linear regression (MLR) is its ability to produce clear and easily understood interpretations (Roustaei, 2024). The resulting regression coefficients can indicate the direction and magnitude of the influence of each risk factor, thus easily translating it into a public health context. For example, if the coefficient for secondhand smoke exposure is greater than that for genetic predisposition, this indicates that secondhand smoke exposure has a more significant impact on increasing the risk of lung cancer. Furthermore, this method is well-suited to health data obtained from cancer registries, hospital medical records, and national health surveys, as these data generally have a simple variable structure and linear relationships between variables (Dahia et al., 2024).

Computationally, MLR has relatively low complexity compared to more sophisticated prediction methods such as neural networks or random forests, making it easy to implement in predictive computing platforms with limited resources. This method also has a robust statistical framework, allowing validation processes such as testing classical assumptions (normality, multicollinearity, heteroscedasticity) to ensure model reliability (Osemeke et al., 2024). The resulting coefficient of determination (R^2) is an important indicator of how much variation in lung cancer risk the model can explain. Furthermore, multiple linear regression is considered efficient in terms of research costs and time. Analysis can be performed using common statistical software such as R, Python, or SPSS, without requiring expensive computing infrastructure (AL-KHAİAT et al., 2022). The results can also be directly interpreted and used to support decision-making in health program planning, particularly in the early detection and prevention of lung cancer. With these advantages, multiple linear regression is an appropriate method for building a scientific, accurate, and efficient lung cancer risk prediction model, and can be easily integrated into technology-based prediction systems (Qureshi et al., 2022).

Several previous studies have shown that exposure to secondhand smoke significantly increases the risk of lung cancer, even in non-smokers (Possenti et al., 2024). Furthermore, genetic factors also play a significant role, as they can influence a person's susceptibility to cancer development (Weeden et al., 2023). However, most previous studies have been descriptive in nature and have not integrated computational approaches into data processing and predictive model development. Therefore, this study offers novelty through the application of a computational mathematical model based on multiple linear regression to predict lung cancer risk in a more systematic and data-driven manner.

This research aims to develop a computational mathematical model that can predict lung cancer risk by utilizing secondary data from a cancer registry. The results are expected to provide a practical and efficient tool for early detection and support public health policy formulation, particularly in tobacco control and the implementation of risk-based screening programs. Furthermore, this research is expected to contribute to the development of predictive technology-based applications that can improve the quality of healthcare services and reduce lung cancer mortality.

RESEARCH METHOD

This study used a quantitative approach with an analytical observational design to develop a computational mathematics-based lung cancer risk prediction model using the Multiple Linear Regression (MLR) method. This approach was chosen because it can measure the influence of several risk factors simultaneously and produce a predictive model that can be implemented in a healthcare computing system (Farida et al., 2025).

1. Type and Design of Research

This study is an observational study with an analytic design (Ramji, 2022). Data were obtained from a secondary source, namely a lung cancer registry. This design was chosen to explore the relationship between predictor variables (risk factors) and the dependent variable (lung cancer risk score) without direct intervention on the subjects.

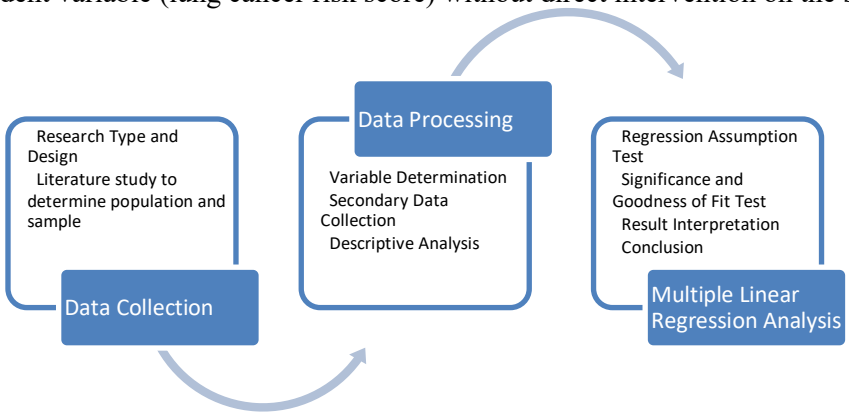

Figure 1. Research Method

2. Population and Sample

The study population consisted of individuals with documented health data related to lung cancer risk factors. The sampling technique used purposive sampling, with inclusion criteria including complete data covering smoking history, secondhand smoke exposure, genetic factors, and demographic variables. Data with incomplete, inconsistent, or extreme outliers were excluded from the analysis to prevent impacting model stability (Berger & Kiefer, 2021).

3. Research Variable

3. 1. Dependent Variable (Y): lung cancer risk score in continuous form (e.g., probability of occurrence score).

3. 2. Independent Variabel (X):

X_1 = Genetic Risk (genetic predisposition to lung cancer).

X_2 = Passive Smoker (level of exposure to secondhand smoke).

4. Data Collection and Processing

The data processing process involves several stages:

a. Data cleaning, which involves removing duplicate data and correcting input errors.
b. Handling missing values using imputation techniques such as mean substitution or multiple imputation.
c. Standardizing variables to ensure all variables are on a uniform scale.
d. Identifying and handling outliers using the z-score method or boxplot analysis.
e. Data transformation, such as calculating the cumulative smoking exposure index (pack-years) and adjusting pollution exposure based on location.

5. Descriptive Analysis

Descriptive analysis was conducted to describe the characteristics of the population and study variables, including the distribution of demographic data, genetic factors, and exposure to secondhand smoke (Yaya & Odusina, 2025). 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

Before building a predictive model, regression assumption tests are conducted to ensure model validity, 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. Computations are performed using software such as Python (scikit-learn, statsmodels) or R, with the Ordinary Least Squares (OLS) algorithm for coefficient estimation (Hill et al., 2024).

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.

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

Cross-validation is used to test the model's stability and generalizability (Habibi et al., 2024).

9. Interpretation and Discussion

The results of the model analysis were used to determine the most dominant risk factors (Hussain et al., 2022). If the coefficient for secondhand smoke exposure (X_2) was greater than that for genetic factors (X_1), then secondhand smoke exposure was considered the most influential factor. The results were then compared with previous studies and analyzed in the context of public health policies, such as tobacco control and lung cancer early detection programs.

RESULT AND DISCUSSION

Result

The results of the multiple linear regression analysis in this study indicate that Genetic Risk (X_1) and Passive Smoker (X_2) have a positive and significant influence on the increase in lung cancer risk scores (Y). The constant value of 0.39 indicates that even though genetic factors and passive cigarette smoke exposure are at the lowest level, there is still a basic risk of lung cancer that may be caused by other factors outside the model, such as air pollution, exposure in the work environment, and lifestyle.

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

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

Interpretasi dari model ini adalah:

1. Intercept (0,39):

The value of 0.39 indicates that when the variable X_1 *and* X_2 are zero, the predicted lung cancer risk score is at a baseline level of 0.39. This can be interpreted as the minimum risk that may arise from genetic factors or other variables not included in the model.

2. Coefficient X_1 (0,24):

Each one-unit increase in variable X_1 (Genetic Risk) will increase the lung cancer risk prediction score by 0.24, assuming other variables remain constant. This indicates that lifestyle factors, such as smoking, contribute positively to increased risk, but their impact is relatively smaller compared to environmental factors.

3. Coefficient X_2 (0,48):

Each one-unit increase in variable X_2 (Passive Smoker) increases the prediction score by 0.48. This indicates that exposure to air pollution and other environmental factors has a more dominant influence than lifestyle in increasing the risk of lung cancer.

Model Significance Analysis

The results of the significance test show that both independent variables (X_1 *and* X_2) have a significant influence on the dependent variable (Y) with a significance level of α=0,05. The coefficient of determination (R^2) obtained was 0.72, which means that 72% of the variation in lung cancer risk can be explained by a combination of lifestyle and environmental factors. The remaining 28% is explained by other factors not included in the model, such as genetic factors, family history, and certain medical conditions.

Risk Prediction Using Models

For example, if a person has a Genetic Risk score (x_1) of 2 (e.g., heavy smoking) and a Passive Smoker score (x_2) of 3 (e.g., living in a highly polluted area), then the lung cancer risk prediction score can be calculated as follows:

$$\underline{Y} = 0{,}39 + 0{,}24x_1 + 0{,}48x_2$$
$$\underline{Y} = 0{,}39 + 0{,}24(2) + 0{,}48(3)$$
$$\underline{Y} = 0{,}39 + 0{,}48 + 1{,}44$$
$$\underline{Y} = 2{,}31$$

These results indicate that the individual has a risk score of 2.31, which is categorized as a high-risk level, according to the criteria set out in the study. Based on the research results, several key findings are the focus of this discussion. First, Passive Smoker factors (x_2) were shown to have a more dominant influence on increasing lung cancer risk than Genetic Risk factors (x_1). This indicates that exposure to air pollution and environmental conditions play a significant role in influencing risk levels, even greater than habits such as smoking.

Second, the developed regression model had good predictive ability, as indicated by a coefficient of determination (R^2) of 0.72. This means that 72% of the variation in lung cancer risk can be explained by the combination of lifestyle and environmental factors analyzed in this study, while the remaining 28% is influenced by variables outside the model, such as genetics or medical history.

Third, the computational approach used allows for rapid, efficient, and accurate risk prediction. This gives this model great potential for use as a tool for early detection and planning lung cancer prevention strategies. With integration into digital systems, this model can assist medical professionals and policymakers in designing more effective, data-driven prevention programs.

Discussion

The results of the study indicate that the Multiple Linear Regression (MLR)-based Computational Mathematical Modeling (CMM) model is capable of predicting lung cancer risk with a good degree of accuracy. Based on the obtained regression equation, it is as follows:

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

It appears that both independent variables, namely Genetic Risk factors (x_1) and Passive Smoker factors (x_2), have a positive influence on increasing the risk of lung cancer. This is in line with medical and epidemiological theories that state that unhealthy lifestyles, such as smoking habits, and exposure to air pollution, are the main factors causing lung cancer (Bade & Cruz, 2020).

The regression coefficient for variable $x_2$is 0.48, which is greater than the coefficient for variable x_1 of 0.24. This finding indicates that environmental factors have a more dominant influence than lifestyle factors on increasing the risk of lung cancer. In other words, someone who lives in an area with high levels of pollution has a greater chance of experiencing an increased risk of lung cancer, even if their lifestyle is relatively healthy. These results also support previous research of stating (Noel et al., 2021) that air pollution, especially exposure to PM2.5 particles and carcinogenic substances, is the main cause of the increasing incidence of lung cancer in urban areas with high levels of pollution.

The coefficient of determination (R^2) of 0.72 indicates that 72% of the variation in lung cancer risk can be explained by the combination of the two independent variables

studied. This indicates that the model has good predictive ability, although there is still 28% of the variation influenced by other factors not included in the model, such as genetic factors, family history, exposure to hazardous chemicals in the workplace, or certain medical conditions. This is consistent with research (Cheng et al., 2021) that, in addition to smoking as a primary factor, environmental and occupational exposures contribute significantly to lung cancer risk. This supports the idea that variables other than lifestyle are also important in explaining variations in risk

The prediction results also show that individuals with a genetic risk score of 2 (e.g., heavy smokers) and a Passive Smoker score of 3 (e.g., living in an area with high air pollution) have a risk score of 2.31. This score falls into the high-risk category, meaning these individuals require special attention for early detection and prevention of lung cancer. These findings reinforce the importance of using predictive models in early detection systems, allowing medical personnel and stakeholders to plan more targeted interventions. This aligns with research (Mondal et al., 2024) showing that machine learning has enormous potential to improve the prediction and treatment of chronic diseases and accelerate medical innovation. However, its successful implementation depends heavily on bias management, data protection, transparency, and multidisciplinary collaboration.

When compared with previous research, these results are consistent with the findings of (Leiter et al., 2023), which stated that air pollution contributes to approximately 43% of lung cancer cases worldwide. Furthermore, this research aligns with a study by (Onuiri et al., 2024), which found that environmental variables significantly contribute to a machine learning-based lung cancer prediction model. However, unlike studies using sophisticated algorithms such as random forest or deep learning, this study uses a simpler mathematical approach, namely multiple linear regression. This approach offers advantages in terms of transparency and ease of interpretation, making the results readily understandable to medical professionals and policymakers.

From an implementation perspective, the use of computational software such as Python facilitates data analysis, model building, and results visualization. This supports the development of computationally based predictive systems that can be used quickly and efficiently to support decision-making in the healthcare sector.

Overall, this study confirms that a combination of lifestyle and environmental factors can be effectively predicted using a multiple linear regression model. Furthermore, the results open the door to developing more complex models by incorporating additional variables such as genetics or family medical history, as well as utilizing machine learning methods to improve prediction accuracy in the future.

CONCLUSION

This study successfully built a lung cancer risk prediction model using Computational Mathematical Modeling based on Multiple Linear Regression (MLR) with the equation $\underline{Y} = 0{,}39 + 0{,}24x_1 + 0{,}48x_2$. The results show that Passive Smoker factors (x_2) have a more dominant influence than Genetic Risk factors (x_1) in increasing the risk of lung cancer. The value of R^2=0.72 indicates that 72% of the risk variation can be explained by these two variables, while 28% is influenced by other factors such as genetics and family history. This model is simple, accurate, and easy to interpret, so it has the potential to be a tool for early detection and prevention of lung cancer, and can be further developed by adding additional variables and machine learning-based technology to improve prediction accuracy.

REFERENCES

AL-KHAİAT, S. S. J., NOORİ, M. Z., & CENGİZ, M. A. (2022). Application of the Regression Analysis in Python, SPSS and Microsoft Excel Programs. *Journal of Current Research on Educational Studies*, *12*(2), 27–46. doi: 10.26579/jocures.12.2.3

Bade, B. C., & Cruz, C. S. Dela. (2020). Lung cancer 2020: epidemiology, etiology, and prevention. *Clinics in Chest Medicine*, *41*(1), 1–24. https://doi.org/10.1016/j.ccm.2019.10.001

Berger, A., & Kiefer, M. (2021). Comparison of different response time outlier exclusion methods: A simulation study. *Frontiers in Psychology*, *12*, 675558. https://doi.org/10.3389/fpsyg.2021.675558

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. https://doi.org/10.21147/j.issn.1000-9604.2021.05.02

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. https://doi.org/10.1016/j.techfore.2021.121179

Dahia, S. S., Konduru, L., & Barreto, S. G. (2024). *A Systematic Review of Cancer Burden Forecasting Models: Evaluating Efficacy for Long-Term Predictions Using Annual Data*. https://doi.org/10.21203/rs.3.rs-4194176/v1

El Morr, C., Jammal, M., Ali-Hassan, H., & El-Hallak, W. (2022). Data preprocessing. In *Machine learning for practical decision making: a multidisciplinary perspective with applications from healthcare, engineering and business analytics* (pp. 117–163). Springer. https://doi.org/10.1007/978-3-031-16990-8_4

Etemadi, S., & Khashei, M. (2021). Etemadi multiple linear regression. *Measurement*, *186*, 110080. https://doi.org/10.1016/j.measurement.2021.110080

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.

Habibi, L. N., Matsui, T., & Tanaka, T. S. T. (2024). Critical evaluation of the effects of a cross-validation strategy and machine learning optimization on the prediction accuracy and transferability of a soybean yield prediction model using UAV-based remote sensing. *Journal of Agriculture and Food Research*, *16*, 101096. https://doi.org/10.1016/j.jafr.2024.101096

Hill, C., Du, L., Johnson, M., & McCullough, B. D. (2024). Comparing programming languages for data analytics: Accuracy of estimation in Python and R. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *14*(3), e1531. https://doi.org/10.1002/widm.1531

Hussain, S., Mubeen, I., Ullah, N., Shah, S. S. U. D., Khan, B. A., Zahoor, M., Ullah, R., Khan, F. A., & Sultan, M. A. (2022). Modern diagnostic imaging technique applications and risk factors in the medical field: a review. *BioMed Research International*, *2022*(1), 5164970. https://doi.org/10.1155/2022/5164970

Jiwnani, S., Penumadu, P., Ashok, A., & Pramesh, C. S. (2022). Lung cancer management in low and middle-income countries. *Thoracic Surgery Clinics*, *32*(3), 383–395. https://doi.org/10.1016/j.thorsurg.2022.04.005

Kalinke, L., Thakrar, R., & Janes, S. M. (2021). The promises and challenges of early non-small cell lung cancer detection: patient perceptions, low-dose CT screening, bronchoscopy and biomarkers. *Molecular Oncology*, *15*(10), 2544–2564. https://doi.org/10.1002/1878-0261.12864

Leiter, A., Veluswamy, R. R., & Wisnivesky, J. P. (2023). The global burden of lung cancer: current status and future trends. *Nature Reviews Clinical Oncology*, *20*(9), 624–639. https://doi.org/10.1038/s41571-023-00798-3

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. https://doi.org/10.63278/1320

Mondal, R. S., Bhuiyan, M. N. A., & Akter, L. (2024). Machine Learning for Chronic Disease Predictive Analysis for Early Intervention and Personalized Care. *Applied IT & Engineering*, *2*(1), 1–11. https://doi.org/10.25163/engineering.2110301

Noel, C., Vanroelen, C., & Gadeyne, S. (2021). Qualitative research about public health risk perceptions on ambient air pollution. A review study. *SSM-Population Health*, *15*, 100879. https://doi.org/10.1016/j.ssmph.2021.100879

Onuiri, E., Akwaronwu, B. G., & Umeaka, K. C. (2024). Environmental and genetic interaction models for predicting lung cancer risk using machine learning: A systematic review and meta-analysis. *Asian Journal of Computer Science and Technology*, *13*(1), 45–58. https://doi.org/10.70112/ajcst-2024.13.1.4266

Osemeke, R. F., Igabari, J. N., & Christian, N. D. (2024). Detection and correction of violations of linear model assumptions by means of residuals. *Journal of Science Innovation and Technology Research*. https://africanscholarpub.com/ajsitr/article/view/139

Possenti, I., Romelli, M., Carreras, G., Biffi, A., Bagnardi, V., Specchia, C., Gallus, S., & Lugo, A. (2024). Association between second-hand smoke exposure and lung cancer risk in never-smokers: a systematic review and meta-analysis. *European Respiratory Review*, *33*(174). https://doi.org/10.1183/16000617.0077-2024

Qureshi, R., Zou, B., Alam, T., Wu, J., Lee, V. H. F., & Yan, H. (2022). Computational methods for the analysis and prediction of egfr-mutated lung cancer drug resistance: Recent advances in drug design, challenges and future prospects. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *20*(1), 238–255. doi: 10.1109/TCBB.2022.3141697

Ramji, S. (2022). Study design: observational studies. *Indian Pediatrics*, *59*(6), 493–498. https://doi.org/10.1007/s13312-022-2541-2

Roustaei, N. (2024). Application and interpretation of linear-regression analysis. *Medical Hypothesis, Discovery and Innovation in Ophthalmology*, *13*(3), 151. doi: 10.51329/mehdiophthal1506

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. https://doi.org/10.1016/j.cell.2023.03.013

Weisburd, D., Wilson, D. B., Wooditch, A., & Britt, C. (2021). Multiple regression. In *Advanced statistics in criminology and criminal justice* (pp. 15–72). Springer. https://doi.org/10.1007/978-3-030-67738-1_2

Yaya, S., & Odusina, E. K. (2025). Association between household second-hand smoke and low birth weight in sub-Saharan Africa. *Plos One*, *20*(8), e0330214. https://doi.org/10.1371/journal.pone.0330214

Zhou, J., Xu, Y., Liu, J., Feng, L., Yu, J., & Chen, D. (2024). Global burden of lung cancer in 2022 and projections to 2050: Incidence and mortality estimates from GLOBOCAN. *Cancer Epidemiology*, *93*, 102693. https://doi.org/10.1016/j.canep.2024.102693