Contemporary Mathematics
http://ojs.wiserpub.com/index.php/CM/

Research Article

Spatio-Temporal Analysis and Prediction by Logistic Regression of Respiratory Diseases in India

Priyanka Subramani, Kalpanapriya Dhakshnamoorthy*

Department of Mathematics, School of Advanced Sciences, Vellore Institute of Technology, Vellore, 632014, Tamil Nadu, India
E-mail: dkalpanapriya@vit.ac.in

Received: 16 October 2024; **Revised:** 11 December 2024; **Accepted:** 2 January 2025

Abstract: Respiratory illnesses rank among the top causes of death and disability in India, influenced by factors such as limited healthcare access, air pollution, smoking, allergens, and a lack of awareness. Despite government efforts to improve respiratory health policies, increase awareness, enhance healthcare facilities, and promote preventive measures, the incidence of respiratory diseases has been on the rise in recent years. This study uses Geographic Information System (GIS) technology to analyze the spatial and temporal distribution patterns of respiratory diseases, aiming to improve our understanding of the contributing factors. Principal component extraction and spatial statistical analyses were utilized to identify the main respiratory illnesses and their geographical distribution. The study concentrated on three major respiratory diseases Tuberculosis, Pneumonia, and Acute Respiratory Distress Syndrome (ARDS) which are related to each other diseases. The findings shows significant variations in the geographical distribution of these diseases across the time period 2019-2021. This spatio-temporal data is essential for enhancing current prevention, control, and treatment strategies for respiratory illnesses in the study area. The methodology applied in this study can be adapted to other regions with similar geographical characteristics and patient data. The study investigated the association of 14 variables with respiratory illnesses. The results indicate that certain variables are associated with an increased risk of frequent flare-ups and hospital admissions due to respiratory diseases. Furthermore, the severity of flare-ups leading to hospital admissions is significantly linked to the presence of comorbidities. These critical and easily measurable variables provide valuable insights for the optimal management of ambulatory patients with respiratory diseases.

***Keywords*:** respiratory diseases, spatio-temporal, logistic regression, risk factors

MSC: 65L05, 34K06, 34K28

Abbreviation

ARDS	Acute Respiratory Disease Syndrome
TB	Tuberculosis
LTBI	Latent Tuberculosis Infection
MDR-TB	Multidrug-Resistant Tuberculosis
GIS	Geographic Information System
PM	Particulate Matter

DOI: https://doi.org/10.37256/cm.6120255923

SD	Standard Deviation
CV	Coefficient of Variation
MLE	Maximum Likelihood Estimation
LRT	Likelihood Ratio Test
EC	Expected Cases
OC	Observed Cases
LLR	Loglikelihood Ratio
RR	Relative Risk
SE	Standard Error
OR	Odds Ratio
CI	Confidence Interval
AIC	Akaike Information Criterion
SBIC	Schwarz Bayesian Information Criterion

1. Introduction

Respiratory diseases encompass a broad spectrum of disorders affecting the lungs and other parts of the respiratory system. These diseases can range from acute infections to chronic conditions, each with distinct pathophysiological mechanisms and clinical manifestations. Among the most significant respiratory diseases are tuberculosis (TB), pneumonia, and acute respiratory distress syndrome (ARDS). These conditions not only pose substantial health challenges but also have profound socio-economic impacts globally. Tuberculosis is a chronic bacterial infection caused by Mycobacterium tuberculosis. It primarily affects the lungs but can spread to other organs. TB remains one of the top ten causes of death worldwide and is particularly prevalent in low and middle-income countries. The transmission of TB occurs through airborne particles, making it highly contagious. The disease manifests in two forms: latent TB infection (LTBI) and active TB disease. LTBI is asymptomatic, while active TB presents with symptoms such as persistent cough, fever, night sweats, and weight loss. The treatment of TB involves a long-term regimen of antibiotics, and the emergence of multidrug-resistant TB (MDR-TB) has made management increasingly complex. Pneumonia is an acute infection that inflames the air sacs in one or both lungs, which may fill with fluid or pus. The condition can be caused by a variety of pathogens, including bacteria, viruses, and fungi. Streptococcus pneumoniae is the most common bacterial cause of pneumonia. Symptoms include cough, fever, chills, and difficulty breathing. Pneumonia can range in severity from mild to life-threatening, particularly in infants, elderly individuals, and those with weakened immune systems. Treatment depends on the underlying cause but often includes antibiotics for bacterial pneumonia and supportive care for viral infections. Acute Respiratory Distress Syndrome is a severe inflammatory condition characterized by rapid onset of widespread inflammation in the lungs. ARDS can result from various direct or indirect injuries to the lung, such as pneumonia, sepsis, trauma, or inhalation injury. The hallmark of ARDS is non-cardiogenic pulmonary edema, leading to severe hypoxemia and respiratory failure. Patients with ARDS typically require mechanical ventilation in intensive care units (ICUs). Despite advances in supportive care, ARDS has a high mortality rate, and survivors often face long-term respiratory and functional impairments. Studies indicate that TB patients have a 39.5% annual risk of developing bacterial pneumonia [1]. Furthermore, about 31% of patients with severe pneumonia may progress to ARDS, depending on the severity and treatment. Data specific to TB leading to ARDS is less common 14%, but severe cases, particularly those involving TB, have been documented to result in ARDS which was shown below in Figure 1.

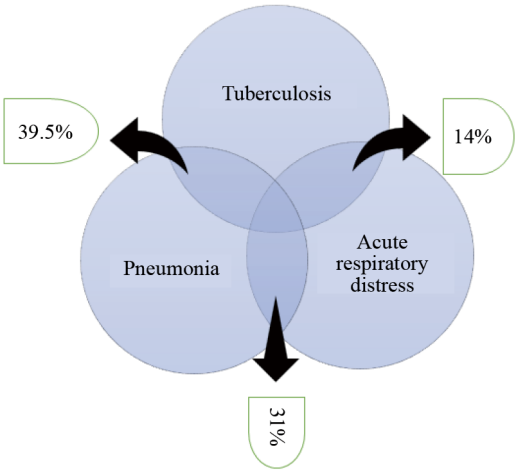


Figure 1. Visualizing relationships of respiratory diseases

Spatial and spatio-temporal data are often encountered in nature, sparking significant interest among scientists across various disciplines, including ecology, epidemiology, and image analysis. Modeling binary data is especially important for understanding events like disease outbreaks or deaths. Spatial and spatio-temporal binary data models are particularly effective for investigating the spread and interaction of a known disease within a grid, where such interactions between neighboring regions are suspected. In spatial epidemiology, spatial clustering analysis is vital for identifying the aggregation of disease cases in specific geographic regions. This analysis determines whether these groupings occur by chance or are statistically significant, providing insights into underlying etiologic factors. Studies have shown that infectious disease distribution is often influenced by various social processes related to their location [2, 3]. The complex interplay of determinants such as socioeconomic vulnerability, rapid population growth, urbanization, and environmental factors can lead to spatial and spatiotemporal variations in respiratory diseases. Identifying and analyzing clusters of respiratory disease, and understanding the factors driving these clusters, is important for investigating outbreaks. Maps generated from these spatial analyses can help prevent and control diseases by enabling targeted public health interventions in areas with elevated disease risk. Additionally, these maps can support public health programs by providing advanced knowledge of disease etiological factors, thereby motivating the population. While spatial analysis focuses solely on spatial variations, spatiotemporal analysis examines how patterns evolve over both space and time, identifying disease trends across spatial units over time. This approach incorporates both spatial and spatiotemporal structures. The spatial resolution of the data plays a crucial role in determining cluster patterns and relevant associations, alongside the types of models used. Different spatial resolutions can produce varying results from the same dataset, regardless of the true extent of spatial correlation. Effects seen at global or regional scales might not be observed at local or individual scales, potentially resulting in an ecological fallacy.

2. Literature review

Anjali et al. [4] identified economic factors as the primary drivers behind the rise in suicide rates in certain regions. Using spatial analysis with a specialized Poisson model, the authors highlighted Madurai as a significant hotspot for various suicide-related factors. To validate their findings, they applied Welch's test, which confirmed Madurai's consistent status as a prominent hotspot. In mainland China, Bie [5] analyzed the influence of seven factors on the relative risk of tuberculosis (TB) through a spatio-temporal distribution model and the INLA algorithm. Meanwhile, Das et al. [6] reported that the increasing elderly population in Singapore serves as the leading risk factor for the recent rise in TB cases, particularly in the southern and eastern regions of the country. Ghazvini [7] developed a logistic regression model incorporating serum markers and body mass index (BMI) to predict TB prevalence with acceptable specificity

and sensitivity. Hoffner [8] emphasized the critical role of Geographic Information Systems (GIS) in identifying high-risk areas and populations vulnerable to TB transmission. Using GIS, the authors studied the relationship between climate and TB distribution in Khuzestan Province, Iran. Similarly, Niu [9] demonstrated the application of logistic regression in educational research to evaluate student performance, considering variables such as study habits and family background. Maja [10] underscored the importance of analyzing specific bio-demographic, socio-economic, and health-related factors to address TB in South Africa. Manish [11] proposed that enhancing literacy and promoting gender equality could help reduce the incidence of rape cases. Marshall [12] provided a comprehensive overview of statistical methodologies for analyzing spatial patterns. Pathak examined the relationship between various variables and TB prevalence among respiratory patients using logistic regression. Ogunsakin [13] investigated the connection between indicators of complications from pulmonary TB and associated risk factors, employing logistic regression. Poonam et al. [14] identified hotspots of crimes against women in Rajasthan using scan statistics. Sukhija [15] highlighted the importance of analyzing real-time, multidimensional data in the Indian state of Haryana. The study pinpointed key hotspots for rape cases in 2017 and compared different hotspot mapping techniques to assess their accuracy in predicting future spatial crime patterns. Subsequent studies by Anjali et al. [16, 17] focused on suicide hotspots and prediction models, revealing alarming trends in India. They utilized time series modeling approaches and demonstrated the superiority of the multivariate VARMA model over the ARIMA model in analyzing factors contributing to suicide hotspots. Zhang [18, 19] emphasized the importance of rigorous feature selection in medical decision-making, particularly for TB prevention strategies, considering meteorological influences. Collectively, these studies provide valuable insights into the complex dynamics of TB and suicide, offering guidance for targeted prevention and intervention efforts. Priyanka et al. [20] identified the top ten states in India with the highest concentration of ARDS cases. Through modeling, they recommended strategies for controlling the contributing factors, which could help mitigate the growth of ARDS infections. Razavi et al. [21] optimized the parameters of a support vector regression (SVR) model for spatio-temporal modeling of asthma-prone areas in Tehran, Iran. Poonam [22] proposed an ensemble hybrid machine learning model for a crime-against-women index, incorporating multiple influencing factors.

After conducting an extensive literature review, we found that spatiotemporal analysis is frequently employed in epidemiology. However, there is limited research on spatiotemporal analysis and prediction using logistic regression in India, particularly concerning specific causes. To address this gap, this study aims to achieve the following goals.

Identifying disease clusters overtime and across different regions to understand the spatial and temporal distribution patterns.

Evaluate the influence of various risk factors on the incidence and spread of respiratory diseases.

Develop predictive models using logistic regression to forecast potential outbreaks or increased incidence of respiratory diseases on historical data and identified risk factors.

Improve the efficiency of disease surveillance systems by integrating spatio-temporal analysis and predictive modeling to detect early warning signs of respiratory disease outbreaks.

3. Data and research design

For our research on respiratory diseases in India, we are utilizing epidemiological data sourced from Indiastat and Nikshay in the period 2019-2021. Indiastat is a comprehensive database that offers a wide range of statistical data across various sectors, including health. It provides detailed information on disease prevalence, demographic data, and health indicators, making it an invaluable resource for understanding broader epidemiological trends of respiratory diseases in India. Nikshay, developed by the Government of India, is a specialized digital platform for tracking and managing tuberculosis (TB) cases, a major respiratory disease in the country. It includes patient-specific data, treatment outcomes, and programmatic indicators, offering granular insights into the incidence, management, and control efforts related to TB [23, 24].

In terms of data design, our research involves integrating and analyzing these datasets to identify patterns and trends in respiratory diseases. Indiastat provides a macro-level perspective with aggregate statistics that help in understanding

the overall burden and distribution of respiratory diseases across different regions and populations. In contrast, the detailed patient-level data from Nikshay facilitates a micro-level analysis, allowing us to examine specific cases, treatment efficacy, and program performance. Combining these data sources enables a comprehensive spatio-temporal analysis, assessing both the broad epidemiological landscape and the detailed, localized impacts of respiratory diseases in India. This integrated approach enhances the robustness of our findings and supports more informed public health interventions. By leveraging data from Indiastat and Nikshay, we can perform a robust spatio-temporal analysis, identifying broad patterns and correlations such as geographic hotspots and the impact of socio-economic factors, while also understanding the specifics of TB cases, treatment efficacy, and patient outcomes. Advanced statistical and geospatial tools, including Geographic Information Systems (GIS), R, Python, and SaTScan, will be employed to visualize data, detect clusters, and model disease spread and intervention impacts.

4. Methodology

4.1 *Spatio-temporal analysis*

Spatio-temporal analysis is a critical field of study that examines how phenomena change across both spatial (geographical) and temporal (time) dimensions. By integrating spatial data, such as geographic coordinates and locations, with temporal data, such as timestamps and time-series, researchers can uncover patterns, trends, and relationships that are not apparent when considering either dimension alone. This type of analysis is employed in various disciplines, including environmental science, where it helps track climate change and natural disasters; urban planning, for monitoring land use and transportation systems; and epidemiology, to trace the spread of diseases and identify health trends. The outline of this analysis is shown in Figure 2. Tools like Geographic Information Systems (GIS), remote sensing, and statistical software such as R, Python, saTScan are commonly used to process and visualize spatio-temporal data. The importance of spatio-temporal analysis lies in its ability to provide deeper insights and support decision-making processes in complex scenarios. For example, in urban planning, understanding traffic patterns over time and space can lead to more efficient transportation systems and reduced congestion. In public health, mapping the spread of an infectious disease can guide effective intervention strategies. However, this field also faces challenges, including the high computational demands of processing large datasets and the difficulty of integrating diverse data sources. Despite these challenges, advances in big data analytics, machine learning, and interdisciplinary collaborations promise to enhance the capabilities and applications of spatio-temporal analysis, making it an indispensable tool in addressing contemporary global issues.

SaTScan, developed by Kulldorff [25, 26], is a specialized software tool designed for the spatial, temporal, and space-time scan statistics analysis. It is widely used in public health, epidemiology, and related fields to detect and evaluate clusters of events, such as disease outbreaks, over time and across geographical regions. The primary function of saTScan is to identify statistically significant clusters by comparing observed data with an expected distribution under the null hypothesis, which assumes no clustering. The software supports various statistical models, including Poisson, Bernoulli, and Normal distributions, allowing for flexibility depending on the nature of the data. SaTScan's scan statistics can be applied to purely spatial analysis, purely temporal analysis, or combined space-time analysis. It helps in pinpointing locations and time periods with unusually high or low event rates, facilitating early detection of outbreaks, environmental hazards, and other critical public health issues. The tool is valued for its robustness, ability to handle large datasets, and its capacity to provide visual and statistical outputs that are essential for effective decision-making and intervention strategies.

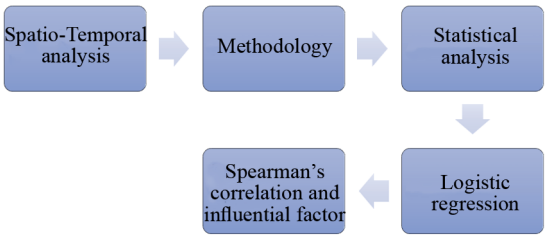


Figure 2. Outline of the work performed

4.2 *Exploratory and descriptive analysis*

Exploratory and descriptive analysis are foundational approaches in data analysis that help researchers understand the basic features of their data. Exploratory Data Analysis (EDA) focuses on discovering patterns, anomalies, and relationships within the data without making any prior assumptions. It involves using statistical summaries, visualizations such as histograms, scatter plots, and box plots, and techniques like clustering to identify underlying structures. EDA is essential for forming hypotheses and guiding subsequent, more formal analyses by providing a comprehensive initial understanding of the dataset.

Descriptive Analysis, on the other hand, aims to summarize and describe the main features of a dataset quantitatively. It provides a straightforward depiction of data through measures of central tendency (mean, median, mode) and measures of variability (range, variance, standard deviation). In addition, descriptive analysis often includes the creation of graphs and tables to effectively communicate the characteristics of the data. This type of analysis is crucial for understanding the general trends and distributions within a dataset, serving as a preliminary step before more complex inferential analyses. Together, exploratory and descriptive analyses form the basis for sound data-driven decision-making and research.

4.3 *Logistic regression*

Logistic regression is a statistical method used for analyzing datasets in which the outcome variable is binary (i.e., it has two possible outcomes) [27]. It is a type of regression analysis that models the probability of a certain class or event, such as presence/absence, success/failure, or yes/no outcomes, based on one or more predictor variables. The logistic regression model estimates the probability that a given instance belongs to a particular category. This is achieved by fitting the data to a logistic function, also known as the sigmoid function, which outputs a value between 0 and 1. The formula for the logistic function is:

$$P(Y=1)=\frac{1}{1+e^{-(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k)}} \tag{1}$$

where $P(Y=1)$ is the probability of the outcome, β_0 is the intercept, $\beta_1,\ \beta_2,\ \cdots,\ \beta_k$ are the coefficients for the predictor variables $X_1,\ X_2,\ \cdots,\ X_k$.

The coefficients β represent the change in the log odds of the outcome for a one-unit increase in the predictor variable. Positive coefficients increase the log odds of the outcome, while negative coefficients decrease them. This method is widely employed in fields such as medicine, social sciences, and marketing due to its ability to handle binary outcomes effectively and provide probabilities and odds ratios, which are straightforward to interpret. Logistic regression can also accommodate multiple predictor variables, making it a versatile and powerful tool for modeling complex relationships [28]. Multiple logistic regression extends simple logistic regression by allowing the inclusion of several predictor variables. This technique is used when the outcome variable is binary, and the goal is to understand the relationship between the outcome and multiple independent variables simultaneously. The exponential of a coefficient e^{β_i} gives the odds ratio for a one-unit change in the predictor variable, offering an intuitive measure of how much more (or less) likely the outcome

is as the predictor variable increases by one unit. The goodness-of-fit for multiple logistic regression can be evaluated using measures such as the Likelihood Ratio Test and the Wald Test [29].

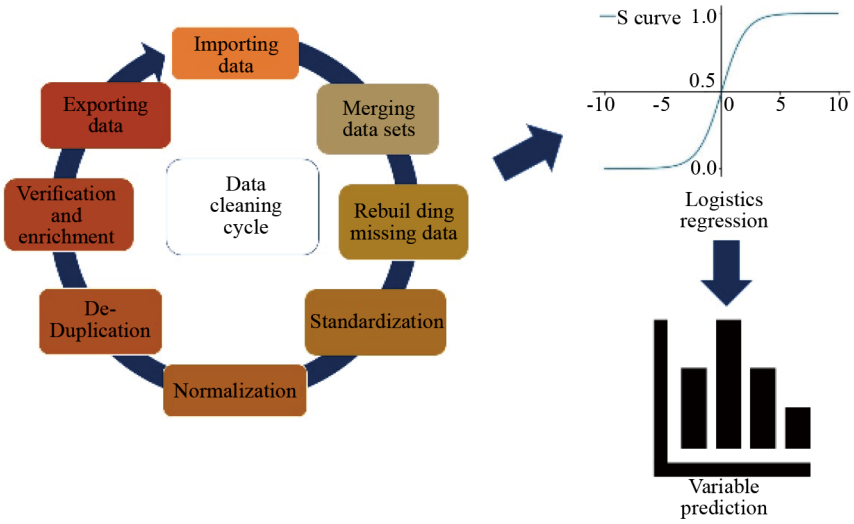


Figure 3. Navigating logistic regression

The Figure 3 provides a concise explanation of the logistic regression process, starting with data cleaning, where raw data is preprocessed to ensure accuracy and consistency by addressing missing values, duplicates, and other inconsistencies. Once the dataset is prepared, logistic regression is applied to model the relationship between independent variables and the dependent variable, calculating probabilities to predict outcomes. Finally, the model uses these probabilities to make predictions about the dependent variable, offering insights into the likelihood of specific outcomes based on the input data.

4.4 *Maximum likelihood estimation*

The maximum likelihood (ML) method stands as the predominant approach for parameter estimation in linear regression models. It is similarly applied to estimate the parameters within logistic regression models. This method, known as maximum likelihood estimation, determines the parameter values of the model that yield the highest likelihood function value. The likelihood function of the model is given below as the formula:

$$L(\beta_i,\ y|x) = p_i^{y_i}(1-p_i)^{1-y_i},\ \text{where } i = 1,\ \cdots k. \tag{2}$$

The log-likelihood function is given below as

$$l(\beta_i,\ y|x) = lnL(\beta_i,\ y|x) = \sum_{y_i=1} ln(p_i) + \sum_{y_i=0} ln(1-p_i) \tag{3}$$

The Maximum Likelihood Estimation (MLE) entails finding the parameter β_i value that maximizes the log-likelihood function 3. This is achieved by solving a specific equation using the Newton-Raphson technique. Here β is the parameter of the logistic model to be estimated, y_i be the dichotomous response variable of the model having the probability p_i and x be the independent variable which possesses the condition on y.

$$\frac{\partial l(\beta_i,\ y|x)}{\partial \beta_i} = 0 \tag{4}$$

4.5 *Assessing the model's performance*

Multiple assessments are conducted to determine the utility, user-friendliness, and adequacy of the model under consideration. The significance of individual variables is examined through coefficient testing, followed by an evaluation of the model's overall fit.

4.5.1 *Wald test*

The Wald test is a statistical hypothesis test used to assess the significance of individual coefficients (parameters) in a regression model, including logistic regression. It evaluates whether a particular predictor variable has a statistically significant effect on the outcome variable.

Formulate Hypothesis: Null Hypothesis (H_0): The coefficient of the predictor variable is equal to zero, indicating no effect on the outcome. Alternative Hypothesis (H_1): The coefficient of the predictor variable is not equal to zero, indicating a significant effect on the outcome.

The test statistic follows as asymptotic chi-square distribution with one degree of freedom under the null hypothesis. The formula for the Wald Test statisti is given below:

$$W = \frac{(\hat{\beta} - \beta_0)^2}{Var(\hat{\beta})} \tag{5}$$

where $\hat{\beta}$ is the estimated coefficient, β_0 is the hypothesized value under the null hypotheis, Var($\hat{\beta}$) is the estimated variance of the coefficient. Here, the Wald test is used to assess the significance of predictor variables in a logistic regression model. It provides a quantitative measure of the impact of each predictor variable on the outcome, helping researchers determine which variables are most influential. Reporting the results of the Wald test enhances the transparency and credibility of the statistical analysis, enabling readers to evaluate the robustness of the findings.

4.5.2 *Likelihood ratio test*

The likelihood ratio test (LRT) is a statistical hypothesis test used to compare the fit of two nested models, typically in the context of logistic regression. In logistic regression, the LRT is commonly employed to evaluate whether adding additional predictor variables significantly improves the fit of the model.

Formulate Hypothesis: Null Hypothesis (H_0'): The simpler model (with fewer predictors) is sufficient to explain the data. Alternative Hypothesis (H_1'): The more complex model (with additional predictors) provides a significantly better fit to the data.

The test statistic follows a chi-square distribution with degrees of freedom equal to the difference in the number of parameters between the two models. The formula for the LRT statistic is given below:

$$LRT = -2 * [(logL_0) - (logL_1)] \tag{6}$$

where L_0 is the parameter with zero and L_1 is the parameter estimated by MLE. The LRT is used to assess the significance of adding new predictor variables to the model. By comparing the fit of nested models, researchers can determine whether the inclusion of additional predictors enhances the model's explanatory power. Reporting the results of the likelihood

ratio test helps to justify the model's complexity and provides insights into the relationship between the predictors and the outcome variable.

4.5.3 *Omnibus test*

The omnibus test in logistic regression is a statistical method used to assess the overall significance of the model as a whole. It examines whether the model, with all its predictor variables collectively, provides a better fit to the data compared to a model with no predictors. In essence, the omnibus test evaluates whether there is a relationship between the predictors and the outcome variable.

Formulate Hypothesis: Null Hypothesis (H_0''): The model with no predictors (null model) fits the data as well as the model with predictors. Alternative Hypothesis (H_1''): The model with predictors provides a significantly better fit to the data than the null model.

The test statistic is typically based on the likelihood ratio or Wald statistic, comparing the fit of the model with predictors to the fit of the null model. The test statistic follows a chi-square distribution with degrees of freedom equal to the difference in the number of parameters between the two models. This is used to evaluate the overall significance of the model. It helps researchers determine whether the predictors collectively contribute to explaining the variation in the outcome variable. Reporting the results of the omnibus test enhances the credibility of the statistical analysis and provides important insights into the overall effectiveness of the logistic regression model in addressing the research question.

Adequacy assessment of the model: The goodness of fit in a statistical model refers to how accurately the model aligns with the observed data and characterizes the dependent variable. Assessing the goodness of fit entails examining the degree of proximity between predicted values and actual observations.

4.5.4 *Hosmer-Lemeshow test*

The Hosmer-Lemeshow test is employed as a statistical approach for evaluating the adequacy of logistic regression models, especially within binary classification scenarios. Its primary function is to gauge the alignment between observed outcomes and the predicted probabilities derived from the logistic regression model. By scrutinizing the model-predicted probabilities correspond to actual outcomes, this test effectively assesses the model's fit. This assessment, in turn, offers valuable insights into the model's performance in prediction and its accuracy in classifying observations. The utilization of the test in reporting findings not only bolsters the credibility of the statistical analysis but also furnishes essential information for deciphering the logistic regression model's soundness and dependability. The Hosmer Lemeshow statistical measure that evaluates the logistic regression model's goodness of fit and accepts any number of independent variables, either qualitative or quantitative. It establishes the significance of the variations between the observed and predicted proportions. Similar to an χ^2 goodness of fit test, the Hosmer-Lemeshow test has the benefit of separating the observations into groups of about similar size, which reduces the likelihood of having groups with extremely low frequencies of observed and predicted values. The anticipated probabilities are used to divide the observations into deciles. Hosmer-Lemeshow statistics are distributed according to χ^2 degree of freedom (D-2) which is given by the below equation 7.

$$\chi^2_{HL} = \sum_{i=1}^{I} \frac{(O_i - E_i)^2}{N_i \varsigma_i (1 - \varsigma_i)} \tag{7}$$

where I is the number of the groups, O_i is the observed events, E_i is the expected events, N_i is the total number of observed events, ς_i is the estimated risk for the i^{th} group.

5. Results and discussions

5.1 *Spatio-temporal of hotspot detection*

In this study, spatio-temporal hotspots of respiratory diseases from 2019 to 2021 were identified using relevant variables through the application of the purely spatio-temporal Poisson model in SaTScan software. The analysis revealed distinct primary and secondary hotspots for each disease, with the results visually represented on a geographical map of India (Figure 4) generated using SaTScan's built-in functions (Google Earth and Cartesian mapping). Notably, Lakshadweep and Kerala consistently emerged as primary hotspots for ARDS, while Rajasthan was identified for Pneumonia, and Delhi for TB across all studied parameters from 2019 to 2021. The detailed results of hotspot detection for the study period are summarized in Tables 1-3. These hotspots have proven to be significant over both the short and long term, suggesting that the government and relevant organizations should prioritize these regions in their policy formulation to address and eradicate these diseases.

Table 1. Hotspot dectection of ARDS by Spatio-temporal analysis

Hotspot	Temporal zone	EC	OC	RR	LLR	*p*-value
Lakshadweep, Kerala	2019-2020 (P)	1,428,529.10	5.62	6.11	7,532,190.59	< 0.00000000000000001
Chandigarh, Himachal pradesh	2019-2020 (S_1)	346,680.84	8.69	8.98	3,892,535.08	< 0.00000000000000001
Rajasthan	2019-2020 (S_2)	3,229,085.51	2.78	3.00	3,656,674.87	< 0.00000000000000001
Gujarat	2019 (S_3)	1,448,066.04	1.68	1.70	282,519.33	< 0.00000000000000001
Sikkim	2019 (S_4)	13,710.56	7.13	7.14	108,069.21	< 0.00000000000000001

Table 2. Hotspot detection of pneumonia by spatio-temporal analysis

Hotspot	Temporal zone	EC	OC	RR	LLR	*p*-value
Rajasthan	2019-2020 (P)	65,131.79	3.39	3.76	122,052.42	< 0.00000000000000001
Chandigarh	2019-2020 (S_1)	989.54	42.89	43.98	118,602.03	< 0.00000000000000001
West bengal	2019-2020 (S_2)	79,648.48	2.66	2.90	80,885.01	< 0.00000000000000001
Uttar pradesh, Haryana	2019 (S_3)	119,967.54	1.77	1.88	31,748.42	< 0.00000000000000001
Andhra pradesh	2021 (S_4)	21,335.45	3.10	3.19	30,759.34	< 0.00000000000000001

Table 3. Hotspot detection of tuberculosis by spatio-temporal analysis

Hotspot	Temporal zone	EC	OC	RR	LLR	*p*-value
Delhi	2019-2020 (P)	65,113.77	2.99	3.05	84,757.78	< 0.00000000000000001
Madhya pradesh	2019 (S_1)	492,920.01	1.31	1.34	23,045.22	< 0.00000000000000001
Andhra pradesh	2019 (S_2)	80,913.26	1.22	1.22	1,864.47	< 0.00000000000000001
Puducherry	2019 (S_3)	2,505.52	1.85	1.85	722.66	< 0.00000000000000001
Nagaland	2019 (S_4)	3,399.04	1.43	1.43	271.89	< 0.00000000000000001

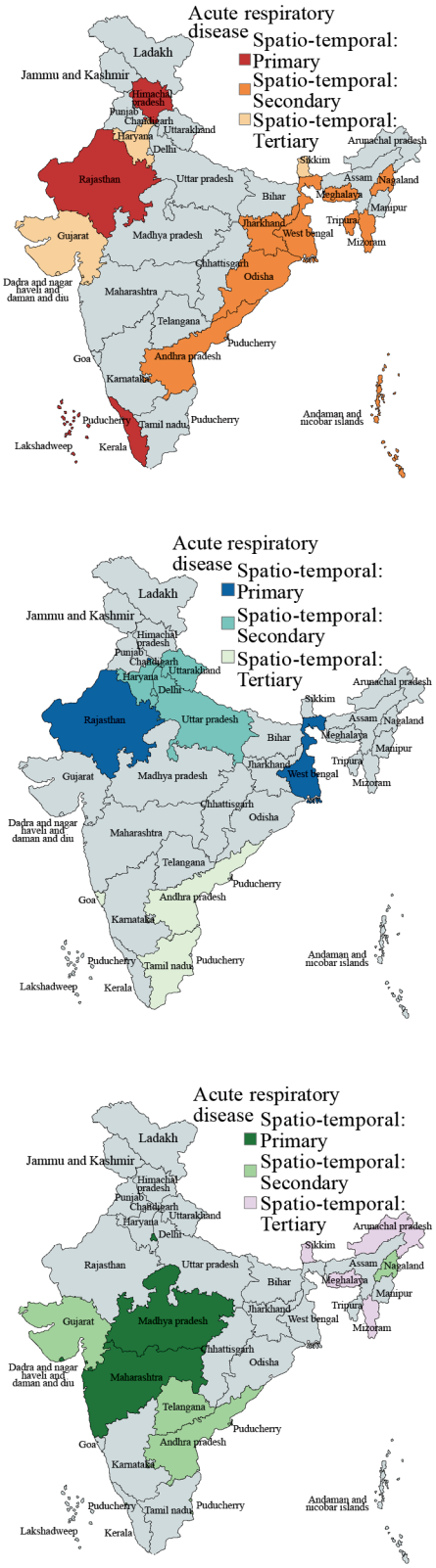


Figure 4. Hotspots of respiratory diseases

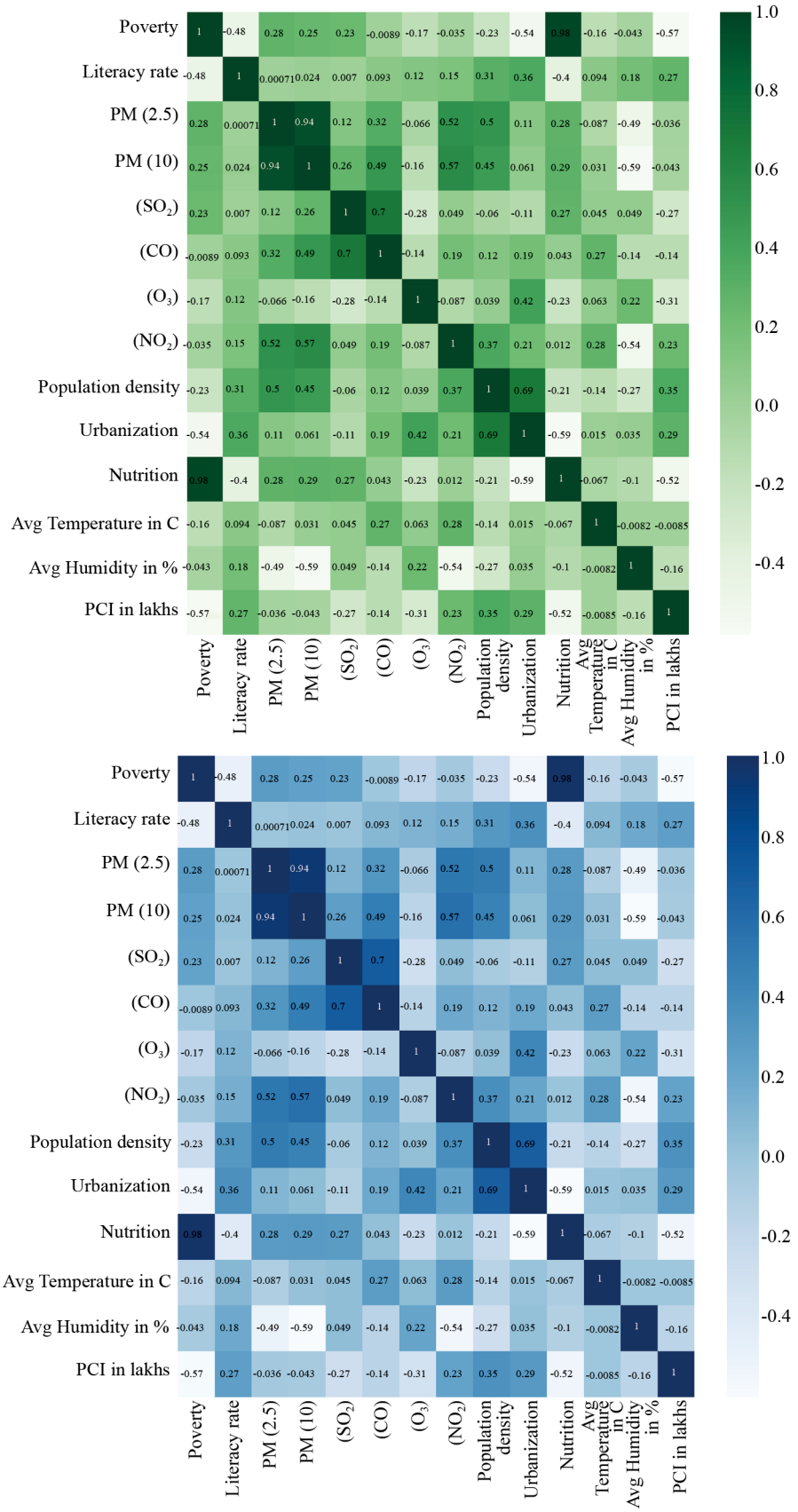
5.2 *Statistical analysis*

Table 4 provides a separate summary of respiratory diseases, offering insights into the data by highlighting the arithmetic mean, standard deviation, sample variance, kurtosis, and skewness. The summary reveals a substantial disparity between the minimum and maximum number of cases over the time period. A higher standard deviation indicates significant data dispersion, underscoring the uneven distribution of respiratory disease cases across the country. Notably, Pneumonia has the lowest mean in India. The states with the highest and lowest consistency in respiratory disease cases were identified using the coefficient of variation (CV). This aids in identifying areas where systematic measures are needed to eradicate the diseases. Although respiratory diseases occur in other states, their incidence is not consistently at an alert level, nor are they considered a severe threat. The less consistent states, predominantly tribal or sparsely populated areas, have shown high variability in cases, indicating volatility. This suggests that while cases exist, they lack specific roots and appear erratic, making them easier to control and monitor. States with higher population and pollution rates are observed to have more consistent cases of respiratory diseases compared to other states.

Table 4. Descriptive statistics of the respiratory diseases

Descriptive statistics	ARDS	Pneumonia	TB	All
Mean	789,790.53	15,930.35	60,414.8	866,135.68
Standard deviation	1,148,743	30,139.97	85,095.80	1,205,523.82
Sample variance	1.32E + 12	9.08E + 08	7.24E + 09	1.45E + 12
Kurtosis	4.88	10.71	10.35	4.35
Skewness	2.16	3.18	2.82	2.07
Count	105	105	105	105

A correlation matrix illustrates the correlation coefficients between variables of respiratory diseases, helping to identify the primary causes that are highly related to other variables. We computed all pairwise Spearmans correlation coefficients between the aligned variables to obtain the matrix. This resulted in a 14 × 14 symmetric matrix with values ranging from -1 to 1. The same matrix, shown in Figure 5, applies to all three diseases as they are caused by the same variables. The analysis revealed that poverty, air pollutants, malnutrition, and population density are statistically significant factors contributing to these diseases. This provides strong evidence of the rise in respiratory disease cases, highlighting the failure of policies to effectively combat and manage these serious health issues during the study period. Government actions appear to be inadequate in addressing the situation.


Poverty
Literacy rate
PM (2.5)
PM (10)
(SO2)
(CO)
(O3)
(NO2)
Population density
Urbanization
Nutrition
Avg Temperature in C
Avg Humidity in %
PCI in lakhs
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4

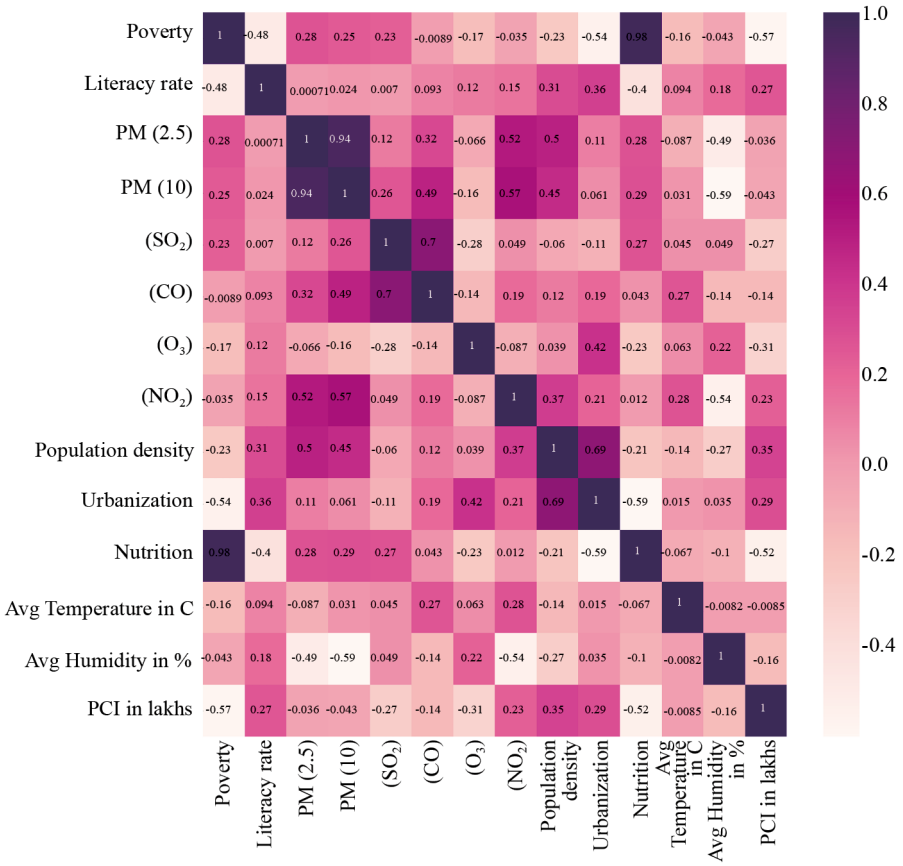


Figure 5. Correlation matrix of the respiratory diseases

5.3 *Logistic regression*

We conducted a multiple logistic regression analysis on the three diseases to understand the significance of various variables and identify related risk factors impacting the dependent variable. Table 5-7 lists the 14 independent variables used to fit the multiple logistic regression model. These tables show the estimated parameters from the fitted model and their Wald Test results. The analysis identified four main factors with significant effects on the diseases: poverty, air pollutants, malnutrition, and temperature. Specifically, PM(2.5), NO_2, O_3 emerged as the most important risk factors among air pollutants. Additionally, for ARDS, literacy rate and population density also had a considerable impact. Malnutrition and Per Capita Income (PCI) played roles in Pneumonia and Tuberculosis cases, though they were less significant compared to the classical risk factors. Overall, air pollutants, poverty, temperature, and population density were identified as critical factors. These findings highlight the importance of considering these variables when studying human characteristics that may influence health outcomes. It is crucial to understand the complex interplay between these factors and their impact on respiratory diseases. These insights have significant implications for respiratory disease research and public health policymakers, aiding in the early detection and prevention of these diseases. The coefficient plot for respiratory diseases is shown separately in Figure 6.

Table 5. Coefficients of multiple logistic regression and wald test for ARDS cases

Variables	Coefficients	SE	z	$P > \|z\|$	χ^2	$P > \chi^2$	Odd ratio
Constant	27.82	16.89	1.64	0.10	2.71	0.09	-
Poverty	-4.49	1.97	-2.27	0.02	5.17	0.02	0.6143
Literacy rate	-0.20	0.15	-1.31	0.18	1.73	0.18	1.3053
PM (2.5)	0.18	0.09	1.96	0.04	3.87	0.04	0.7777
PM (10)	-0.11	0.06	-1.81	0.07	3.28	0.06	0.6636
SO_2	0.53	0.35	1.50	0.13	2.24	0.13	1.7145
CO	-0.001	0.003	-0.50	0.61	0.25	0.61	1.0931
O_3	0.05	0.07	0.71	0.47	0.51	0.47	1.1368
NO_2	0.004	0.09	0.05	0.95	0.002	0.95	1.0622
Population density	-0.001	0.001	-1.15	0.24	1.32	0.24	0.9650
Urbanization	0.15	0.12	1.21	0.22	1.48	0.22	0.8513
Malnutrition	5.11	2.26	2.26	0.02	5.12	0.02	0.7236
Temperature	-0.47	0.26	-1.79	0.07	3.22	0.07	1.2624
Humidity	0.11	0.07	1.57	0.11	2.46	0.11	1.7610
PCI	-2.91	1.50	-1.94	0.05	3.76	0.05	1.0365

Table 6. Coefficients of multiple logistic regression and wald test for pneumonia cases

Variables	Coefficients	SE	z	$P > \|z\|$	χ^2	$P > \chi^2$	Odd ratio
Constant	11.24	9.14	1.53	0.32	1.47	0.07	-
Poverty	-3.21	1.22	-1.12	0.009	3.24	0.01	0.6854
Literacy rate	0.42	0.15	1.34	0.15	1.54	0.12	0.7940
PM (2.5)	0.62	0.06	-0.82	0.15	0.87	0.53	1.5188
PM (10)	0.67	0.04	0.28	0.67	0.08	0.88	1.5972
SO_2	0.009	0.01	0.29	0.86	2.34	0.38	0.7625
CO	0.17	0.02	-0.24	0.78	0.34	0.57	1.0174
O_3	0.92	0.07	0.62	0.39	0.66	0.51	0.8232
NO_2	0.66	0.36	0.05	0.93	0.003	0.88	2.8929
Population density	0.68	0.75	-1.65	0.45	1.56	0.53	1.6627
Urbanization	-0.26	0.32	1.67	0.38	1.13	0.26	1.6923
Malnutrition	0.05	1.14	1.02	0.54	4.89	0.13	0.7507
Temperature	-0.30	0.43	-1.54	0.43	3.21	0.09	1.9155
Humidity	0.24	0.13	1.32	0.13	1.97	0.15	0.5081
PCI	-1.34	1.98	-1.11	0.12	2.65	0.14	2.3268

Table 7. Coefficients of multiple logistic regression and wald test for tuberculosis cases

Variables	Coefficients	SE	z	$P > \|z\|$	χ^2	$P > \chi^2$	Odd ratio
Constant	7.95	14.19	0.56	0.57	0.31	0.57	-
Poverty	1.82	1.39	1.30	0.19	1.70	0.19	0.7540
Literacy rate	-0.21	0.15	-1.45	0.14	2.10	0.14	0.9395
PM (2.5)	-0.05	0.06	-0.97	0.33	0.94	0.33	0.8030
PM (10)	0.01	0.04	0.43	0.66	0.19	0.66	0.9218
SO_2	0.02	0.25	0.10	0.91	0.01	0.91	1.4056
CO	0.004	0.004	1.09	0.27	1.20	0.27	1.5697
O_3	-0.31	0.17	-1.75	0.08	3.06	0.07	0.5771
NO_2	0.23	0.12	1.92	0.05	3.71	0.05	1.2953
Population density	0.0002	0.001	0.27	0.78	0.07	0.78	1.3348
Urbanization	0.02	0.10	0.25	0.79	0.06	0.79	1.4977
Malnutrition	-2.13	1.70	-1.25	0.21	1.56	0.21	0.7574
Temperature	-0.07	0.22	-0.34	0.72	0.12	0.72	1.3428
Humidity	0.08	0.07	1.16	0.24	1.36	0.24	1.5409
PCI	1.20	1.01	1.19	0.23	1.42	0.23	1.6099

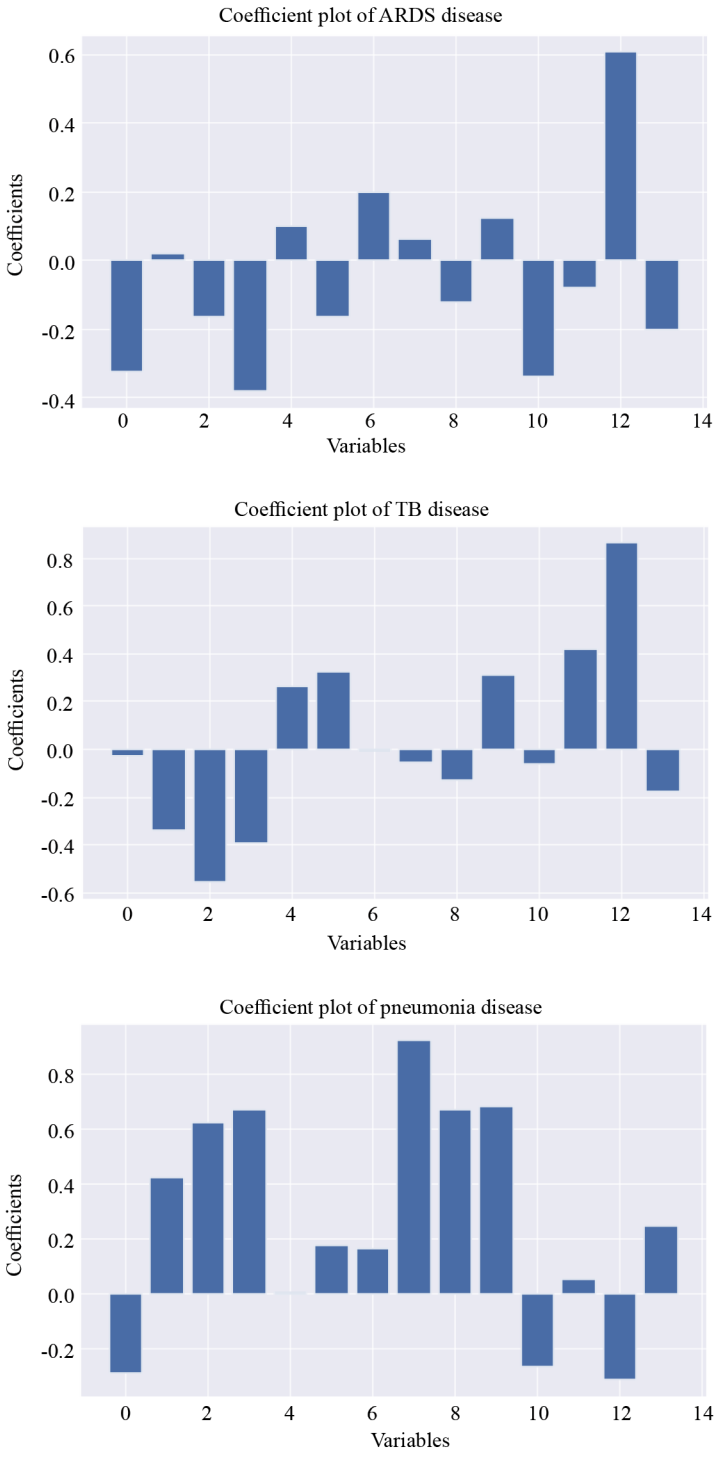


Figure 6. Coefficient plot of the respiratory diseases

After extracting all non-significant factors, we refitted the multiple logistic regression model using only the significant independent variables. The Wald test for the coefficients of these variables indicates their significant contribution to predicting the diseases. The odds ratio (OR) of each significant variable is presented in the table, with

a 95% confidence interval (CI) for comparison with all other variables. Table 8 provides the model selection values, including AIC, BIC, and deviance for the respiratory diseases.

Table 8. Model selection criteria

Respiratory diseases	AIC	BIC	Deviance
ARDS	53.89	47.23	1.0
Pneumonia	53.12	46.78	1.75
Tuberculosis	54.12	47.47	1.25

5.3.1 *Accessing the model performance*

Table 9 presents the likelihood ratio test and omnibus test results for all three diseases, indicating that the independent variables significantly contribute to predicting the main causes of the diseases. Although the constant has no practical interpretation in the model, it is typically retained regardless of its significance. Table 10 displays the contingency table for the Hosmer-Lemeshow test, showing the goodness of fit for each disease at different p-values. Consequently, the overall model fit is good.

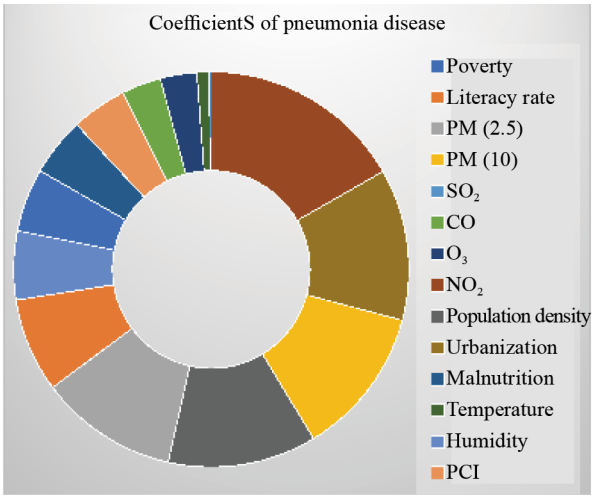
Table 9. Test of model effects

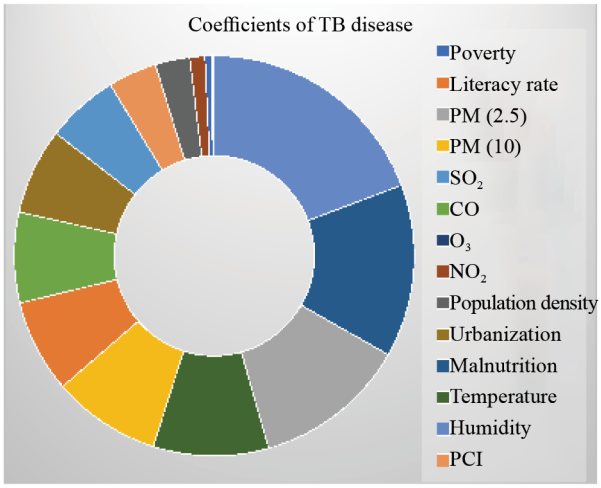
Respiratory diseases	LRT test	Omnibus test
ARDS	$\chi^2:\ 22.49\ p:<0.01$	$\chi^2:\ 26.01\ p:<0.01$
Pneumonia	$\chi^2:\ 22.41\ p:<0.01$	$\chi^2:\ 25.12\ p:<0.01$
TB	$\chi^2:\ 22.90\ p:<0.01$	$\chi^2:\ 25.77\ p:<0.01$

Table 10. Contigency table

Respiratory diseases	Hosmer-Lemeshow test
ARDS	$\chi^2 12.89\ p:\ 0.58$
Pneumonia	$\chi^2 7.89\ p:\ 0.69$
TB	$\chi^2 11.81\ p:\ 0.52$

To evaluate the success of the logistic regression model in predicting respiratory diseases, various performance measures were used, as discussed in the methodology section. These measures include sensitivity, specificity, test F_1 score, train F_1 score, and accuracy. Ideally, all these metrics would equal one; however, this is generally unattainable, especially for respiratory diseases. We compared these measures for ARDS, Pneumonia, and Tuberculosis. Initially, logistic regression was applied to each of the three diseases, using their respective variables as the dependent variable. The objective was to assess the model accuracy in predicting the diseases based on the available variables. The factors affecting the stability of the diseases are shown in the Figure 7. Once the model was successfully calibrated, the goal was to compare the relative importance of each risk factor identified by their respective criteria. It is understood that the causes of the three diseases function similarly, and the prediction metrics of the model are presented in Table 11.





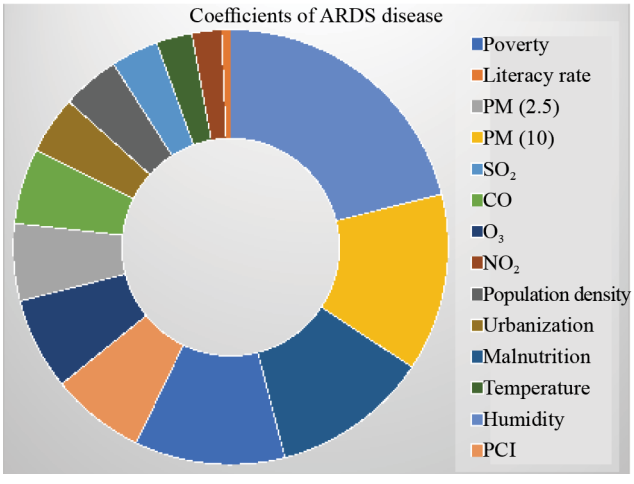


Figure 7. Coefficients of the respiratory diseases

Table 11. Performance of logistic regression

Respiratory diseases	Sensitivity	Specificity	Test F_1 score	Train F_1 score	Accuracy
ARDS	0.22	0.78	0.76	0.82	0.81
Pneumonia	0.32	0.68	0.85	0.88	0.72
TB	0.10	0.90	0.76	0.77	0.92

6. Conclusion

One of the crucial aspects in the medical industry is predicting respiratory diseases using available data. Numerous techniques and methods exist for this purpose. In this research, we identified spatio-temporal hotspots and applied logistic regression. The key element is the selection of data and variables, which enhances the accuracy and effectiveness of the method. Our proposed statistical framework did not reveal a significant association but identified the classical risk factors among the three common respiratory diseases, providing a modeling approach to understand the relative importance of these risk factors in their respective datasets. While air pollutants and malnutrition significantly impact one disease, they may not have the same effect on the other two. These inconsistent results highlight the need for further research to clarify the main causes of respiratory diseases. According to the results of the statistical tests, the multiple logistic regression model has performed efficiently. This study emphasizes the importance of community involvement in limiting the spread and eradicating these diseases as early as possible. People should adhere to preventive policies and increase societal awareness and commitment to the precautionary measures recommended by the Ministry of Health.

Acknowledgement

The Vellore Institute of Technology, Vellore, has been acknowledged by the authors for giving us the resources we needed to complete this research project effectively.

Funding

There was no particular grant awarded for this research by governmental, private, or non-profit funding organizations.

Conflict of interest

The authors declare no competing financial interest.

References

[1] Lin CH, Lin CJ, Kuo YW, Wang JY, Hsu CL, Chen JM, et al. Tuberculosis mortality: Patient characteristics and causes. *BMC Infectious Diseases*. 2014; 14: 1-8.

[2] Pathak AK, Sharma M, Katiyar SK, Katiyar S, Nagar PK. Logistic regression analysis of environmental and other variables and incidences of tuberculosis in respiratory patients. *Scientific Reports*. 2020; 10(1): 21843.

[3] Rosli NM, Shah SA, Mahmood MI. Geographical Information System (GIS) application in tuberculosis spatial clustering studies: A systematic review. *Malaysian Journal of Public Health Medicine*. 2018; 18(1): 70-80.

[4] Kumar BR. Spatial analysis of multivariate factors influencing suicide hotspots in Urban Tamil Nadu. *Journal of Affective Disorders Reports*. 2024; 16: 100741.

[5] Bie S, Hu X, Zhang H, Wang K, Dou Z. Influential factors and spatial-temporal distribution of tuberculosis in mainland China. *Scientific Reports*. 2021; 11(1): 6274.
[6] Das S, Cook AR, Wah W, Win KMK, Chee CBE, Wang YT, et al. Spatial dynamics of TB within a highly urbanised Asian metropolis using point patterns. *Scientific Reports*. 2017; 7(1): 36.
[7] Ghazvini K, Mansouri S, Shakeri MT, Youssefi M, Derakhshan M, Keikha M. Prediction of tuberculosis using a logistic regression model. *Reviews in Clinical Medicine*. 2019; 6(3): 108.
[8] Hoffner S, Hadadi M, Rajaei E, Farnia P, Ahmadi M, Jaberansari Z, et al. Geographic characterization of the tuberculosis epidemiology in iran using a geographical information system. *Biomedical and Biotechnology Research Journal*. 2018; 2(3): 213-219.
[9] Niu L. A review of the application of logistic regression in educational research: Common issues, implications, and suggestions. *Educational Review*. 2020; 72(1): 41-67.
[10] Maja TF, Maposa D. An investigation of risk factors associated with tuberculosis transmission in South Africa using logistic regression model. *Infectious Disease Reports*. 2022; 14(4): 609-620.
[11] Manish GV, Simran, Kumar J, Choubey DK. Identification of hotspot of rape cases in NCT of delhi: A data science perspective. In: *International Conference on Information Systems and Management Science*. Springer; 2021. p.485-496.
[12] Marshall RJ. A review of methods for the statistical analysis of spatial patterns of disease. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. 1991; 154(3): 421-441.
[13] Ogunsakin R, Adebayo A. Performance of logistic regression in tuberculosis data. *International Journal of Scientific and Research Publications*. 2014; 4(9): 1.
[14] Saravag PK. An application of scan statistics in identification and analysis of hotspot of crime against women in rajasthan, india. *Applied Spatial Analysis and Policy*. 2024; 17: 1-20.
[15] Sukhija K, Singh SN, Kumar J. Spatial visualization approach for detecting criminal hotspots: An analysis of total cognizable crimes in the state of Haryana. In: *2017 2nd IEEE International Conference on Recent Trends in Electronics, Information and Communication Technology (RTEICT)*. IEEE; 2017. p.1060-1066.
[16] Anjali B, Kumar J. Spatio-temporal aspect of suicide and suicidal ideation: An application of SaTScan to detect hotspots in four major cities of tamil nadu. *Journal of Scientific Research*. 2021; 65(9): 7-18.
[17] Kumar R. Exploring cause-specific strategies for suicide prevention in India: A multivariate VARMA approach. *Asian Journal of Psychiatry*. 2024; 92: 103871.
[18] Zhang Z, Trevino V, Hoseini SS, Belciug S, Boopathi AM, Zhang P, et al. Variable selection in logistic regression model with genetic algorithm. *Annals of Translational Medicine*. 2018; 6(3): 45.
[19] Zhang Y, Liu M, Wu SS, Jiang H, Zhang J, Wang S, et al. Spatial distribution of tuberculosis and its association with meteorological factors in mainland China. *BMC Infectious Diseases*. 2019; 19: 1-7.
[20] Subramani P, Dhakshnamoorthy K. Spatial aspects of acute respiratory disease syndrome: An application of scan statistics using satscan in identification and analysis of hotspot in india. *Contemporary Mathematics*. 2024; 5(3): 3804-3821.
[21] Razavi-Termeh SV, Sadeghi-Niaraki A, Choi SM. Spatio-temporal modelling of asthma-prone areas using a machine learning optimized with metaheuristic algorithms. *Geocarto International*. 2022; 37(25): 9917-9942.
[22] Saravag PK, Kumar BR. *A Hybrid Machine Learning and Regression Approach for Validating a Multi-Dimensional Crime Index in the Context of Crime against Women*. IEEE; 2024.
[23] Li Z, Liu Q, Zhan M, Tao B, Wang J, Lu W. Meteorological factors contribute to the risk of pulmonary tuberculosis: A multicenter study in eastern China. *Science of The Total Environment*. 2021; 793: 148621.
[24] Kiezun A, Lee ITA, Shomron N. Evaluation of optimization techniques for variable selection in logistic regression applied to diagnosis of myocardial infarction. *Bioinformation*. 2009; 3(7): 311.
[25] Kulldorff M. A spatial scan statistic. *Communications in Statistics-Theory and Methods*. 1997; 26(6): 1481-1496.
[26] Kulldorff MS. *V8. 0: Software for the Spatial and Space-Time Scan Statistics*. Information Management Services, Inc.; 2009.
[27] Allison P. *Logistic Regression Using SAS: Theory and Application*. SAS Institute; 2012.
[28] Harrell FE. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer; 2001.
[29] Tabachnick BG, Fidell LS, Ullman JB, Fidell LS. *Using Multivariate Statistics*. Allyn and Bacon; 2001.