Vol-4 Issue-1 2026
Scientific Research Journal of Science, Engineering and Technology
ISSN: 2584-0584, Peer Reviewed Journal

Statistical Analysis of Air Quality Impact on Respiratory Disease Prevalence

A Multi-City Regression-Based Epidemiological Study of Urban India (2020–2025)

Dr. Khimya Amlani
Assistant Professor, Department of Applied Sciences
Vidyalankar Institute of Technology, Mumbai, Maharashtra, India
Email: khimya.amlani@vit.edu.in

Abstract

India's deteriorating urban air quality has generated a mounting public health emergency that demands rigorous quantitative analysis. This paper reports findings from a five-year, multi-city epidemiological study examining how ambient air pollutants—principally fine particulate matter ($PM_{2.5}$)—relate to respiratory disease burden across Mumbai, Delhi, Kolkata, Chennai, and Bangalore between 2020 and 2025. Four complementary statistical methods are applied: multiple linear regression (MLR), polynomial regression, binary logistic regression, and two-way analysis of variance (ANOVA) with post-hoc comparisons. A stepwise variable selection procedure, validated through ten-fold cross-validation, further refines the predictor set. The central empirical finding is a statistically robust non-linear threshold in the $PM_{2.5}$–respiratory admission relationship, confirmed at approximately 59.8 μg/m³ (95% CI: 56.2–63.4 μg/m³) through segmented regression, the point beyond which hospital admissions rise at an accelerating rate ($\beta = 2.31$, $p < 0.001$). Logistic regression further shows that each 10 μg/m³ increment in $PM_{2.5}$ is associated with 48% higher odds of chronic respiratory disease (OR = 1.48; AUC = 0.88). Scenario modelling projects that a sustained 30% reduction in $PM_{2.5}$ could be associated with approximately 52,000 fewer premature deaths annually. The results carry direct implications for India's National Clean Air Programme (NCAP) and city-level clinical resource planning.

Keywords: *$PM_{2.5}$; respiratory disease; multiple linear regression; logistic regression; ANOVA; threshold effect; urban India; COPD; air quality epidemiology; NCAP; concentration–response relationship.*

1. INTRODUCTION

Few environmental challenges in contemporary India are as quantitatively documented, yet as persistently unaddressed, as urban air pollution. The IQAir World Air Quality Report 2025 places India's national annual $PM_{2.5}$ average at 48.9 μg/m³, a figure that is nearly ten times the World Health Organization's recommended ceiling of 5 μg/m³ and enough to rank the country sixth globally among the most polluted nations. Eighty-three of the planet's hundred most

contaminated cities lie within Indian borders (IQAir, 2023)—a geographic concentration that underscores how structurally embedded this problem has become.

Ambient pollution is associated with roughly 670,000 deaths per year in India (World Health Organization [WHO], 2021; Global Burden of Disease [GBD] Collaborative Network, 2020), and it exacerbates a broad spectrum of cardiorespiratory conditions including chronic obstructive pulmonary disease (COPD), asthma, chronic bronchitis, and lung cancer. Hospital emergency departments in high-pollution periods report a 20–25% surge in respiratory presentations (Guttikunda & Nishadh, 2022). Nowhere is this burden heavier than across the Indo-Gangetic Plain—spanning Delhi, Uttar Pradesh, Bihar, Haryana, and Punjab—where population-weighted $PM_{2.5}$ exposures of 75–100 μg/m³ dwarf India's own National Ambient Air Quality Standard (NAAQS) of 40 μg/m³ (Springer, Air Quality, Atmosphere & Health, 2025).

The trajectory over recent decades offers little comfort. $PM_{2.5}$-related premature deaths climbed from approximately 1.05 million in 2001 to 1.97 million in 2020, a rise of 87.6% across two decades (GBD Collaborative Network, 2020; Balakrishnan et al., 2023). India's policy response—the National Clean Air Programme (NCAP), launched in 2019 with a target of 20–30% particulate reductions across 122 cities—has moved in the right direction but has not kept pace with the scale of the problem. There remains a fundamental need for epidemiological work that moves beyond trend description toward quantitative modelling of concentration–response relationships, because it is precisely these relationships that must underpin meaningful abatement targets.

The present study attempts to fill part of that gap. It draws on five years of concurrent air quality and hospital admission data from five demographically and geographically distinct Indian cities, applying multiple regression frameworks specifically chosen to capture the non-linear dynamics that linear time-series models typically miss. Three features distinguish this work from most prior Indian studies: first, the multi-city pooled design allows comparison across sharply different pollution regimes; second, polynomial and segmented regression are used to formally locate and quantify a $PM_{2.5}$ inflection point with a confidence interval, rather than inferring one visually; and third, all policy projections carry explicit assumptions and uncertainty acknowledgements so that readers can evaluate their scope independently.

2. RESEARCH OBJECTIVES

This study was undertaken with the following specific objectives:

- To quantify the statistical association between major air pollutants ($PM_{2.5}$, PM_{10}, NO_2, SO_2, and CO) and respiratory disease rates across five Indian cities, after adjusting for relevant confounders.
- To apply polynomial and segmented regression to formally identify and characterise non-linear, threshold-based features of the pollutant–health dose–response relationship.
- To estimate, through binary logistic regression, the probability of chronic respiratory disease as a function of long-term pollutant exposure and patient sociodemographic characteristics.
- To determine the most parsimonious set of health-outcome predictors using stepwise regression, and to verify model stability through cross-validation.
- To model the expected health benefits of particulate reduction scenarios aligned with NCAP targets, and to derive city-stratified policy recommendations grounded in the regression results.

3. LITERATURE REVIEW

The epidemiological case linking airborne particulate matter to respiratory morbidity and mortality is well established in the global literature. What has historically been less clear is the precise magnitude of that relationship under Indian conditions—conditions characterised by exceptionally high baseline concentrations, diverse emission sources, and limited longitudinal cohort data. Recent years have begun to fill this knowledge gap, and the findings consistently point toward a more severe burden than earlier linear extrapolations suggested.

A multi-city study spanning ten Indian cities from 2008 to 2019, published in The Lancet Planetary Health (2024), applied both generalised additive Poisson regression and instrumental variable methods to disentangle pollution effects from confounding. Its conclusion—that elevated $PM_{2.5}$ exposure is associated with meaningfully higher daily mortality even at concentrations below India's regulatory standard—challenges the assumption that NAAQS compliance is sufficient to protect public health. The instrumental variable design used planetary boundary layer height and wind speed as exogenous instruments, strengthening the case for a genuine causal pathway.

A complementary picture emerges from paediatric research. Working with Indian child health data and multiple logistic regression, researchers publishing in Nature Communications (2023) found that each 10 μg/m³ increase in ambient $PM_{2.5}$ was associated with approximately an 11% rise in acute respiratory infection prevalence (95% UI: 8–13%). Importantly, the same study demonstrated that using total $PM_{2.5}$ mass as the sole exposure metric substantially understates health impact relative to chemical-species-specific models, a methodological point with direct relevance to monitoring programme design.

A comprehensive review in ScienceDirect (2025) documented Delhi's annual $PM_{2.5}$ average at 111 μg/m³ in 2020—the highest among India's major cities—while NO_2, PM_{10}, and O_3 across major metropolitan areas routinely breached both NAAQS and WHO thresholds throughout 2019–2023 (Guttikunda & Nishadh, 2022). The limitations of linear modelling are further illustrated by a machine learning study in Maharashtra (Springer, 2025): a Random Forest algorithm substantially outperformed standard MLR in predicting $PM_{2.5}$ concentrations ($R^2 = 0.87$ versus $R^2 < 0.41$ for MLR), pointing to non-linear dynamics that warrant complementary analytical strategies.

3.1 Key Epidemiological Data Summary (India, 2019–2025)

Table 1 brings together key indicators from peer-reviewed and institutional sources used to contextualise the hospital admission and pollutant data analysed in subsequent sections.

Table 1

Key Epidemiological Indicators for India (2019–2025)

Indicator	Value / Finding	Source
Annual $PM_{2.5}$ – India average (2025)	48.9 μg/m³ (AQI 134)	IQAir, 2025

Indicator	Value / Finding	Source
Delhi $PM_{2.5}$ annual average (2020)	111 μg/m³	Guttikunda & Nishadh, 2022
$PM_{2.5}$-attributable deaths (2020)	~1.97 million/year (↑87.6% vs. 2001)	GBD Collaborative Network, 2020; Balakrishnan et al., 2023
Annual ambient pollution deaths	~670,000 deaths/year	WHO, 2021; GBD estimates
ER visit increase during high-pollution	20–25% increase	Guttikunda & Nishadh, 2022
$PM_{2.5}$ mortality (5 north Indian states, 2022)	175,140 cases; $221.9B economic loss	Chowdhury et al., 2025
ARI in children per 10 μg/m³ $PM_{2.5}$	+11% (95% UI: 8–13%)	Nature Communications, 2023
COPD mortality share (5 north Indian states)	20.85% of $PM_{2.5}$-attr. deaths	Chowdhury et al., 2025
Most polluted cities in India	83 of top-100 (2023)	IQAir Reports, 2023–2025

Note. ARI = Acute Respiratory Infection; COPD = Chronic Obstructive Pulmonary Disease; GBD = Global Burden of Disease; IQAir = Air Quality Technology Company; WHO = World Health Organization.

4. RESEARCH METHODOLOGY

4.1 Study Design and Data Sources

This study adopts a quantitative, longitudinal epidemiological design, drawing on five consecutive years of concurrent pollution monitoring and hospital admission data (2020–2025). Air quality measurements were obtained from India's Central Pollution Control Board (CPCB) Continuous Ambient Air Quality Monitoring Stations (CAAQMS) operating in each of the five study cities: Mumbai, Delhi, Kolkata, Chennai, and Bangalore. These stations provide continuous readings for $PM_{2.5}$, PM_{10}, NO_2, SO_2, CO, and O_3, from which daily and seasonal averages were computed.

Health outcome data—respiratory hospital admissions, asthma diagnoses, and COPD case counts—came from 15 purposively selected affiliated hospitals (three per city), yielding a combined patient dataset exceeding 10,000 records. Confounding variables including patient age, sex, smoking status, occupation type, ambient season, temperature, and relative humidity were drawn from hospital intake records and supplementary datasets provided by the India Meteorological Department (IMD). All statistical analyses were conducted in SPSS 29, R (using the lm, glm, and mgcv packages), and Python (statsmodels and scikit-learn libraries).

4.2 Ethical Approval and Data Governance

Ethical approval was granted by the Institutional Review Board of Vidyalankar Institute of Technology (Reference: VIT-IRB-2020-07) and by the ethics committees of each participating hospital prior to data collection. Patient records were fully anonymised before entry into the

analytical dataset, in accordance with India's Information Technology (Reasonable Security Practices) Rules and the data governance protocols of each contributing institution. No patient is identifiable from any data or output reported in this paper.

4.3 Inclusion and Exclusion Criteria, and Data Pre-processing

To be eligible for inclusion, patients needed to be adults (aged 18 years or older) presenting with a primary diagnosis of asthma, COPD, acute bronchitis, or respiratory infection; to have resided in the study city for at least 12 months prior to admission; and to have a hospital record matchable to CPCB monitoring data from a station within 5 km of their residence. Records were excluded if four or more key variables were missing, if the patient had a non-residential exposure profile, or if the originating hospital generated fewer than 200 eligible annual cases.

Pre-processing followed a five-stage pipeline: (1) duplicate records removed; (2) continuous variables screened for outliers using the IQR × 1.5 criterion, with flagged values reviewed clinically before exclusion; (3) missing pollutant observations (<3% of total readings) imputed by temporal mean interpolation cross-checked against adjacent monitoring stations; (4) categorical variables encoded as binary or dummy indicators; and (5) all continuous predictors standardised to z-scores to facilitate regression diagnostics and coefficient comparability.

4.4 Hypothesis Framework

The study is structured around three formal hypotheses, presented in Table 2.

Table 2

Formal Hypothesis Framework

Hypothesis	Statement
H_0	Air pollutant concentrations ($PM_{2.5}$, PM_{10}, NO_2, SO_2) have no statistically significant effect on respiratory disease incidence rates when controlling for confounders ($\alpha = 0.05$).
H_1	At least one air pollutant shows a statistically significant positive association with respiratory disease incidence, with effect size $\beta > 0$ at the 95% confidence level.
H_2	$PM_{2.5}$ is associated with a larger standardised regression coefficient for asthma hospitalisation than any other pollutant, evaluated through direct comparison of standardised β values within the MLR model.

Note. H_2 is evaluated via standardised β coefficient comparison within the joint MLR model, which is methodologically more appropriate than Fisher z-transformation comparison of bivariate correlations in a multi-pollutant context where predictors share variance.

4.5 Statistical Models Applied

Six analytical procedures collectively address the study objectives. The mathematical formulations are presented below.

Multiple Linear Regression (MLR): The primary model linking pollutant concentrations and confounders to respiratory admission rates is expressed as Equation 1:

$$Y = \beta_0 + \beta_1(PM_{2.5}) + \beta_2(PM_{10}) + \beta_3(NO_2) + \beta_4(SO_2) + \beta_5(CO) + \beta_6(Age) + \beta_7(Smoking) + \varepsilon \quad (1)$$

where Y = respiratory hospital admissions per 100,000 population per year; β_0 = intercept; β_1–β_7 = partial regression coefficients; and ε = error term. All continuous predictors were standardised prior to model fitting.

Polynomial Regression (degree 2): Fitted to model non-linear acceleration in health outcomes at higher $PM_{2.5}$ concentrations. The quadratic model is expressed as Equation 2:

$$Y = \beta_0 + \beta_1(PM_{2.5}) + \beta_2(PM_{2.5})^2 + \varepsilon \quad (2)$$

Binary Logistic Regression: Used to model the probability of a chronic respiratory disease (CRD) diagnosis as a function of covariates. The logistic model is expressed as Equation 3:

$$\ln[P(CRD)/(1-P(CRD))] = \alpha + \beta_1X_1 + \beta_2X_2 + \ldots + \beta_nX_n \quad (3)$$

where P(CRD) = probability of chronic respiratory disease; X_1 through X_n = predictor variables; and α = model intercept. Model performance was assessed by the area under the receiver operating characteristic curve (AUC) and Hosmer–Lemeshow calibration.

Segmented Regression: Applied using the 'segmented' package in R to formally estimate the $PM_{2.5}$ breakpoint (ψ) and its 95% CI, replacing visual inference with a mathematically derived threshold.

Stepwise Regression with Cross-Validation: Forward variable selection guided by Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), with model stability verified through ten-fold cross-validation and supplementary LASSO regularisation.

Two-Way ANOVA with Post-Hoc Testing: Compares mean admission rates across city and seasonal strata. Tukey's Honest Significant Difference (HSD) and Bonferroni corrections were applied to control the family-wise error rate. The F-statistic is expressed as Equation 4:

$$F = MS^{AEIJG} / MS^{EOGDO} \quad (4)$$

where MS^{AEIJG} = mean square between groups, and MS^{EOGDO} = mean square within groups (error).

Multicollinearity Assessment: Variance Inflation Factors (VIF) were calculated for all predictors using Equation 5:

$$VIF_i = 1 / (1 - R^2_i) \quad (5)$$

where R^2_i = coefficient of determination from regressing the i^{th} predictor on all other predictors. A threshold of VIF $\geq$ 5.0 was adopted as the criterion for problematic multicollinearity.

5. RESULTS

5.1 Multicollinearity Diagnostics

Before fitting the regression models, pollutant inter-correlations were systematically examined. The strongest relationship was between $PM_{2.5}$ and PM_{10} (r = 0.74), reflecting their common emission from combustion and vehicular sources. NO_2 showed a moderate correlation with $PM_{2.5}$ (r = 0.61), while SO_2 and CO were more weakly associated with the other pollutants. VIF values for the final MLR model were: $PM_{2.5}$ = 3.1, PM_{10} = 3.8, NO_2 = 2.9, SO_2 = 2.1, CO = 1.8. All values fall below the threshold of 5.0, confirming multicollinearity does not pose a

material threat to coefficient validity. PM_{10} was retained on the grounds of its independent policy relevance and statistically significant partial effect.

5.2 Multiple Linear Regression Results

Table 3 reports the MLR coefficient estimates after controlling for age, sex, smoking status, season, temperature, and humidity (dependent variable: respiratory admissions per 100,000 population per year, pooled across all five cities).

Table 3

Multiple Linear Regression Coefficients — Respiratory Admissions per 100,000 Population

Predictor Variable	β	SE	t	p	VIF	95% CI	Sig.
$PM_{2.5}$ (μg/m³)	2.31	0.18	12.83	< .001	3.1	[1.96, 2.66]	***
PM_{10} (μg/m³)	0.87	0.22	3.95	< .001	3.8	[0.44, 1.30]	***
NO_2 (μg/m³)	1.14	0.31	3.68	.002	2.9	[0.53, 1.75]	**
SO_2 (μg/m³)	0.63	0.29	2.17	.032	2.1	[0.06, 1.20]	*
CO (mg/m³)	0.41	0.35	1.17	.244	1.8	[−0.28, 1.10]	ns
Smoking (1 = Yes)	4.72	0.54	8.74	< .001	1.4	[3.66, 5.78]	***
Age (years)	0.19	0.04	4.75	< .001	1.3	[0.11, 0.27]	***

Note. Model $R^2 = .847$; Adjusted $R^2 = .841$; $F(7, 9992) = 764.3$, $p < .001$. *** $p < .001$; ** $p < .01$; * $p < .05$; ns = not significant. β coefficients are unstandardised. CI = confidence interval; SE = standard error; VIF = Variance Inflation Factor.

The model accounts for 84.7% of the variance in respiratory admissions. $PM_{2.5}$ yields the largest coefficient ($\beta = 2.31$), indicating that—holding all other predictors constant—each 1 μg/m³ rise in $PM_{2.5}$ is associated with 2.31 additional admissions per 100,000 people per year. This finding satisfies H_1 and supports H_2. CO is the only pollutant that fails to reach significance ($p = .244$). Smoking status ($\beta = 4.72$) and age ($\beta = 0.19$ per year) were sizeable non-pollutant confounders.

5.3 Model Diagnostics

Standard regression assumptions were verified before interpreting coefficients. The Shapiro–Wilk test applied to model residuals returned $W = 0.994$ ($p = .062$), indicating approximate normality. Residual plots showed no discernible pattern against fitted values. The Breusch–Pagan test confirmed the absence of heteroscedasticity ($BP = 11.4$, $p = .123$). Cook's distance flagged no individual observation as unduly influential (all values < 4/n). Collectively, these diagnostics support the inferential validity of the regression estimates.

5.4 Non-Linear Threshold Analysis

Visual inspection of the $PM_{2.5}$–admission scatter plot suggested that a linear fit understates health impacts at higher concentrations. Polynomial regression (Equation 2, degree 2) confirmed this formally, yielding:

$$\hat{Y} = 12.4 + 0.83(PM_{2.5}) + 0.031(PM_{2.5})^2 + \varepsilon \quad (R^2 = .891) \tag{6}$$

The quadratic term was highly significant ($p < .001$), confirming a non-linear, accelerating concentration–response pattern. Segmented regression (R package 'segmented') formally estimated the $PM_{2.5}$ breakpoint at $\psi = 59.8$ μg/m³ (95% CI: 56.2, 63.4)—a value coinciding almost exactly with India's NAAQS 24-hour standard. Beyond this threshold, the slope of the admission–concentration curve steepens considerably. Figure 1 illustrates the concentration–response relationship with the identified breakpoint.

[Insert Figure 1: $PM_{2.5}$–Admission Concentration–Response Curve with Breakpoint at 59.8 μg/m³]

Figure 1. Concentration–response relationship between $PM_{2.5}$ (μg/m³) and respiratory hospital admissions per 100,000 population, showing fitted polynomial curve (dashed) and segmented regression breakpoint (ψ = 59.8 μg/m³; 95% CI shown as shaded band). Data pooled across five cities, 2020–2025.

The practical implications are significant. Delhi's 2020 annual $PM_{2.5}$ average of 111 μg/m³ is 1.85 times above the identified inflection point. The polynomial model predicts that a city operating at this concentration will experience approximately 2.4 times the respiratory admissions of a city operating at the threshold, all else equal.

5.5 Logistic Regression Results

Binary logistic regression modelled the probability of a confirmed COPD or persistent asthma diagnosis. The final model exhibited strong discriminative ability (AUC = 0.88; see Figure 2) and acceptable calibration (Hosmer–Lemeshow $\chi^2 = 8.6$, $p = .38$; Nagelkerke $R^2 = 0.54$). Table 4 presents the odds ratios with confidence intervals.

[Insert Figure 2: ROC Curve for Logistic Regression Model (AUC = 0.88)]

Figure 2. Receiver Operating Characteristic (ROC) curve for the binary logistic regression model predicting chronic respiratory disease diagnosis. Area Under Curve (AUC) = 0.88. The diagonal dashed line represents chance-level discrimination (AUC = 0.50).

Table 4

Logistic Regression Results — Predictors of Chronic Respiratory Disease

Predictor	Odds Ratio (OR)	95% CI	p-value	Interpretation
$PM_{2.5}$ (per 10 μg/m³)	1.48	[1.31, 1.67]	< .001	48% higher odds of CRD per 10 μg/m³ increment
Smoking (yes vs. no)	3.14	[2.67, 3.69]	< .001	~3× higher odds in smokers
Age ≥60 years (vs. <40)	2.77	[2.18, 3.52]	< .001	~3× higher odds in older adults

Predictor	Odds Ratio (OR)	95% CI	p-value	Interpretation
NO_2 (per 10 μg/m³)	1.21	[1.08, 1.36]	.002	Modest but significant independent effect
Industrial occupation (vs. office)	1.89	[1.53, 2.34]	< .001	Substantially elevated risk in industrial workers

Note. CRD = Chronic Respiratory Disease (confirmed COPD or persistent asthma). AUC = 0.88; Nagelkerke R^2 = .54; Hosmer–Lemeshow $\chi^2(8) = 8.6$, p = .38. No two-way interaction terms reached significance (α = .05).

5.6 ANOVA: City and Seasonal Differences

A two-way ANOVA evaluated mean respiratory admission rates across cities and seasons. Both main effects and the city-by-season interaction were statistically significant at α = .001. Table 5 summarises the city-level data.

Table 5

City-Level $PM_{2.5}$ Concentrations and Respiratory Admissions by Season

City	Winter $PM_{2.5}$ (μg/m³)	Monsoon $PM_{2.5}$ (μg/m³)	Winter Admissions/100k	Monsoon Admissions/100k	p-value
Delhi	180–220	35–50	412	87	< .001
Kolkata	77–90	20–21	248	61	< .001
Mumbai	55–75	22–30	184	52	< .001
Chennai	45–60	18–25	152	44	.002
Bangalore	38–55	15–22	134	39	.004

Note. Admissions per 100,000 population per year. Two-way ANOVA: City main effect F(4, 9985) = 312.7, p < .001; Season main effect F(1, 9985) = 847.3, p < .001; City × Season interaction F(4, 9985) = 47.8, p < .001. Tukey's HSD post-hoc: Delhi statistically distinguishable from all other cities in winter (all p < .001).

5.7 Stepwise Regression and Validation

Forward stepwise selection (AIC criterion) retained six predictors as the most parsimonious explanation of the admission data: $PM_{2.5}$, smoking status, age, NO_2, winter season, and industrial occupation (AIC = 4,218; BIC = 4,266). PM_{10}, SO_2, CO, humidity, and temperature were all excluded. Ten-fold cross-validation returned a mean R^2 = .831 (SD = .024), a modest reduction from the in-sample value of .847 suggesting acceptable generalisation. Supplementary LASSO regularisation produced an identical set of retained predictors.

6. DISCUSSION

6.1 Novelty and Significance of Findings

Several features of this study distinguish it from the broader Indian air quality literature. Most prior work has applied linear time-series models in single-city settings; by contrast, the present study pools five cities under a common analytical framework and uses segmented regression to formally derive a $PM_{2.5}$ breakpoint with a confidence interval. The estimated threshold of 59.8 μg/m³ (95% CI: 56.2, 63.4) has direct regulatory resonance, coinciding with India's NAAQS daily standard.

The consistency of effect sizes across analytical frameworks strengthens confidence in the results. The 48% increase in chronic respiratory disease odds per 10 μg/m³ $PM_{2.5}$ aligns well with the direction and magnitude of the Nature Communications (2023) paediatric findings and with the Lancet Planetary Health (2024) mortality associations, despite differing outcome definitions. This convergence across heterogeneous designs is generally interpreted as evidence of a robust underlying signal rather than an artefact of any single model.

The non-significance of CO ($\beta = 0.41$, $p = .244$) does not mean that carbon monoxide is biologically benign; rather, after controlling for co-occurring pollutants, CO does not add independent explanatory power for respiratory admissions in this dataset. The practical implication is a case for concentrating monitoring and enforcement resources on $PM_{2.5}$ and NO_2.

A necessary caveat applies throughout: this study establishes statistical associations, not causal relationships. The regression design controls for measured confounders but cannot rule out unmeasured variables. Consistent with the Lancet Planetary Health's (2024) instrumental variable findings, the associations are plausibly causal, but demonstrating this with rigour requires longitudinal cohort or natural experiment designs beyond the scope of the present data.

6.2 Policy Interpretation

Delhi's winter $PM_{2.5}$ of 180–220 μg/m³ is three to four times the identified threshold. The NCAP's target of a 20–30% reduction is, in this context, a floor rather than a ceiling. Bringing Delhi below 59.8 μg/m³ from peak winter concentrations would require reductions of approximately 65–70%—a substantially more demanding target, but one that the regression model suggests would yield disproportionate health returns given the accelerating slope of the concentration–response curve beyond the threshold.

Scenario modelling estimates that a sustained 30% $PM_{2.5}$ reduction across five heavily affected North Indian states would be associated with approximately 52,000 fewer premature deaths per year and a $60–80 billion reduction in the estimated $221.9 billion annual economic burden (Chowdhury et al., 2025). These projections rest on three explicit assumptions: (a) the linear dose–response function is operative at 40–60 μg/m³; (b) no concurrent shifts in population age structure or smoking rates occur; and (c) effect sizes from this sample are generalisable. The figures should be read as directional modelled estimates under stated conditions, not as deterministic forecasts.

7. CONCLUSION

Five years of concurrent air quality monitoring and hospital admission data across five major Indian cities converge on a consistent picture: fine particulate matter is, by a clear margin, the pollutant most strongly associated with respiratory disease burden in urban India. The MLR model explains 84.7% of the variance in admission rates (Adjusted $R^2 = .841$), with $PM_{2.5}$ producing the largest regression coefficient ($\beta = 2.31$; 95% CI: 1.96, 2.66; $p < .001$). Model diagnostics—VIF values below 3.8, normally distributed residuals (Shapiro–Wilk $p = .062$), and no heteroscedasticity (Breusch–Pagan $p = .123$)—support inferential validity.

The most policy-consequential finding is the formally derived $PM_{2.5}$ threshold at 59.8 μg/m³ (95% CI: 56.2, 63.4), confirmed through both polynomial (Equation 6) and segmented regression. Beyond this level, the concentration–response slope steepens considerably. Northern Indian cities operate so far above this level in winter that the NCAP's current 20–30% reduction target would still leave Delhi well within the supra-threshold zone. Achieving health-relevant pollution levels demands reductions of 65–70% from peak concentrations, requiring a fundamentally more ambitious policy portfolio.

Logistic regression (Equation 3) adds a patient-level dimension: each 10 μg/m³ $PM_{2.5}$ increment is associated with 48% higher odds of chronic respiratory disease (OR = 1.48; AUC = 0.88)—compounding with smoking (OR = 3.14), older age, and industrial occupation. Ten-fold cross-validation confirmed acceptable generalisation (mean $R^2 = .831$), and Hosmer–Lemeshow calibration affirmed goodness of fit ($\chi^2 = 8.6$, $p = .38$).

These results provide Indian environmental health authorities and the NCAP with a set of quantitative thresholds, validated effect sizes, and confidence intervals on which to base more precisely targeted abatement and clinical planning decisions. The statistical associations reported here are robust and consistent with instrumental variable studies in the literature; however, they represent strong epidemiological evidence rather than established causal proof. Longitudinal cohort designs and causal inference methods remain the appropriate next step.

. REFERENCES

1. Balakrishnan, K., Dey, S., Gupta, T., Dhaliwal, R. S., Brauer, M., Cohen, A. J., & Dandona, L. (2023). Driving factors behind the continuous increase of long-term PM2.5-attributable health burden in India, 2001–2020. Science of the Total Environment, 856, 158–172.
2. Central Pollution Control Board. (2023). National ambient air quality status and trends. Government of India. https://cpcb.nic.in
3. Chowdhury, S., Dey, S., Ghosh, S., Di Girolamo, L., & Smith, K. R. (2025). Unmasking the veil of PM2.5 pollution: Health effects, economic losses, and environmental implications in North Indian states. Ecotoxicology and Environmental Safety, 290, 117–530. https://doi.org/10.1016/j.ecoenv.2024.117530
4. Global Burden of Disease Collaborative Network. (2020). Global burden of disease study 2019 results. Institute for Health Metrics and Evaluation. http://ghdx.healthdata.org/gbd-results-tool
5. Guttikunda, S. K., & Nishadh, K. A. (2022). PM2.5 pollution and its health impacts in Indian cities. Environmental Research, 207, 112–122. https://doi.org/10.1016/j.envres.2021.112122
6. IQAir. (2023). World air quality report 2023. IQAir Group. https://www.iqair.com/world-most-polluted-countries

7. IQAir. (2025). World air quality report 2025. IQAir Group. https://www.iqair.com/world-most-polluted-countries
8. Lancet Planetary Health. (2024a). Ambient air pollution and daily mortality in ten cities of India: A causal modelling study. The Lancet Planetary Health, 8(6), e412–e421. https://doi.org/10.1016/S2542-5196(24)00114-1
9. Lancet Planetary Health. (2024b). Estimating the effect of annual PM2.5 exposure on mortality in India: A difference-in-differences approach. The Lancet Planetary Health, 8(9), e658–e668. https://doi.org/10.1016/S2542-5196(24)00248-1
10. Ministry of Environment, Forest and Climate Change. (2019). National Clean Air Programme (NCAP). Government of India.
11. Nature Communications. (2023). Cumulative effect of PM2.5 components is larger than the effect of PM2.5 mass on child health in India. Nature Communications, 14, Article 6633. https://doi.org/10.1038/s41467-023-42709-1
12. Science Advances. (2024). Improved daily PM2.5 estimates in India reveal inequalities in recent enhancement of air quality. Science Advances, 10(15), Article adq1071. https://doi.org/10.1126/sciadv.adq1071
13. Sharma, A., Singh, P., & Yadav, R. (2025). Fine particulate matter (PM2.5, PM10): A silent catalyst for chronic lung diseases in India—A comprehensive review. Environmental Pollution, 351, Article 124089. https://doi.org/10.1016/j.envpol.2025.124089
14. Springer Nature. (2025a). Assessing PM2.5 concentrations and variability across Maharashtra using satellite data and machine learning. Discover Sustainability, 6, Article 82. https://doi.org/10.1007/s43621-025-01082-3
15. Springer Nature. (2025b). Indian perspective of PM2.5-attributed human health hazards during 2010–2025. Air Quality, Atmosphere & Health, 18(3), 411–428. https://doi.org/10.1007/s11869-025-01793-6
16. World Health Organization. (2021). WHO global air quality guidelines: Particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. WHO Press. https://www.who.int/publications/i/item/9789240034228