Data-Driven Insights into Asthma Management: Analysis of Risk Factors and Outcomes

Course: ALY6015 – Applied Analytics

Instructor: Richard He

Students: A. S. Sushmitha Urs, Lavanya Suresh, Vaibhavi Bograja Raj

Date: 12, December, 2025

Introduction

Asthma is a chronic inflammatory condition of the airways that affects millions of individuals worldwide and remains a major contributor to emergency room visits, preventable hospitalizations, and healthcare burden. Identifying the factors that influence asthma severity, patient risk levels, and patterns of acute care utilization is essential for improving disease management and developing effective intervention strategies. This project analyzes a synthetic asthma dataset consisting of 10,000 patient records and 17 clinically meaningful variables, covering demographic information (age, gender, BMI), lifestyle factors (smoking status, physical activity), environmental exposures (air pollution), clinical biomarkers (FeNO, peak expiratory flow), comorbidities, medication adherence, and emergency room (ER) visit history. The dataset provides a realistic and comprehensive snapshot of patient-level factors commonly examined in asthma research.

Using a combination of exploratory analysis, statistical modeling, and nonparametric validation, this study examines how individual characteristics and clinical measurements relate to asthma outcomes such as disease severity, ER utilization, and diagnostic likelihood. By integrating multiple analytical methods, the project aims to generate data-driven insights that support better clinical decision-making, inform public health strategies, and enhance our understanding of asthma risk patterns within a diverse patient population.

Methods

Data Overview

The synthetic asthma dataset used in this analysis contains 10,000 patient records and includes 17 clinically relevant variables spanning demographics, lifestyle behaviors, environmental exposures, clinical biomarkers, comorbidities, medication adherence, and healthcare utilization outcomes. The dataset provides a comprehensive and realistic foundation for exploring asthma-related patterns across a diverse simulated patient population.

Key Variable Categories

Demographic Variables

- Age: continuous variable ranging from early childhood to late adulthood
- Gender: categorical (Male, Female, Other)
- BMI: continuous measure of body composition

Lifestyle and Behavioral Factors

- Smoking_Status: categorical indicator of tobacco exposure
- Physical_Activity_Level: categorical measure of physical activity intensity

Environmental Exposure

- Air_Pollution_Level: ordered factor (Low, Moderate, High), representing ambient air quality conditions

Clinical Biomarkers and Health Indicators

- FeNO_Level: marker of airway inflammation
- Peak_Expiratory_Flow: measure of lung function
- Family_History: binary indicator of inherited asthma risk
- Comorbidities: categorical representation of chronic conditions such as diabetes or hypertension

Treatment Behavior

- Medication_Adherence: continuous value from 0 to 1 indicating adherence to prescribed asthma therapy

Asthma Outcomes

- Has_Asthma: binary indicator of asthma diagnosis
- Asthma_Control_Level: ordered categorical measure of severity (Well Controlled, Poorly Controlled, Not Controlled)
- Number_of_ER_Visits: count of emergency room visits (0–6), representing healthcare utilization.

Statistical Methods

The dataset was partitioned into training (75%, n=7,500) and testing (25%, n=2,500) subsets for model validation. We employed Pearson correlation analysis to examine bivariate relationships and identify multicollinearity. Multiple linear regression and Poisson regression were applied to model ER visits, with Poisson preferred for count data. Logistic regression with LASSO regularization predicted asthma diagnosis, providing odds ratios and enabling variable selection. Ordinal logistic regression modeled severity as it appropriately handles ordered categorical outcomes. Interaction models tested whether medication adherence effectiveness depends on air pollution level. ANOVA compared mean ER visits across control levels, with Kruskal-Wallis as nonparametric alternative. Additional tests included Wilcoxon rank-sum for group comparisons and Spearman correlation for monotonic relationships.

Model performance was evaluated using confusion matrices, accuracy, AIC, and ROC curves with AUC.

Analysis

Exploratory Data Analysis

R Code:

```r
asthma <- read.csv("synthetic_asthma_dataset.csv", stringsAsFactors = TRUE)
summary(asthma)
table(asthma$Has_Asthma)
```

Table 1: Key Dataset Characteristics

Variable	**Category/Statistic**	**Value**
Total Patients	N	10,000
Has_Asthma	No (0)	7,567 (75.67%)
	Yes (1)	2,433 (24.33%)
Asthma_Control_Level	Well Controlled	84 (3.45%)
	Poorly Controlled	1,120 (46.03%)
	Not Controlled	1,229 (50.49%)
Age	Range (Peak)	1-89 years (40-50)
BMI	Mean (Range)	25.05 (15-45) kg/m²
Medication_Adherence	Mean	0.498 (50%)
ER Visits	Mean (Range)	1.016 (0-6)
High Utilizers	2+ visits	2,684 (26.84%)
Family History	Yes	3,034 (30.34%)

Patient ages showed concentration in middle age with approximately 430 patients at 40-50 years compared to 220-230 at extremes. The gender distribution was balanced (48% female, 48% male, 4% other). Emergency room visits were heavily right-skewed with 70% experiencing zero visits. Among asthma patients, only 3.45% achieved well-controlled status, indicating substantial management challenges.

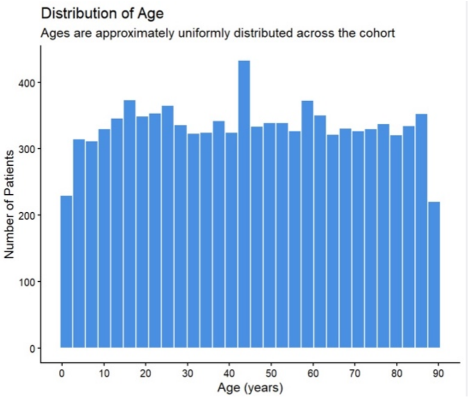


Figure 1: Age Distribution

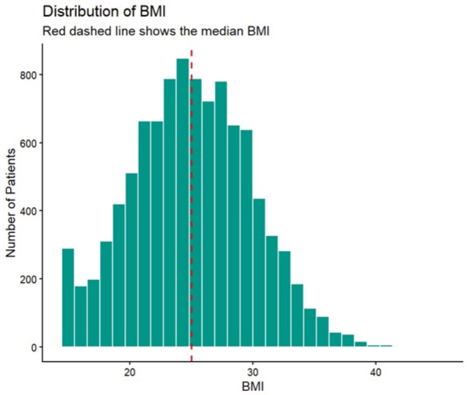


Figure 2: BMI Distribution

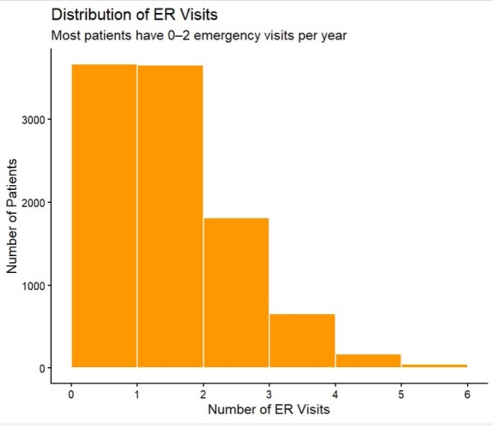


Figure 3: ER Visits Distribution

Correlation Analysis

R Code:

```
r

cor_matrix <- cor(asthma[, num_vars], use = "complete.obs")
```

Correlation analysis revealed Family_History and Has_Asthma as the strongest correlated pair ($r = 0.41$, $p < 0.001$), confirming the genetic component of asthma susceptibility. BMI showed weaker but significant correlation with asthma ($r = 0.10$, $p < 0.001$). Most other numerical variables demonstrated very weak linear associations with emergency room visits (all $|r| < 0.02$), including age ($r = -0.015$), medication adherence ($r = -0.001$), and FeNO level ($r = 0.010$), suggesting relationships may be more complex than simple linear associations. Minimal multicollinearity was observed among predictors (all pairwise $|r| < 0.42$).

Multiple Linear and Poisson Regression

R Code:

```
r

lm_fit <- lm(Number_of_ER_Visits ~ Age + BMI + Family_History +
             Medication_Adherence + FeNO_Level + Has_Asthma, data = asthma)
pois_fit <- glm(Number_of_ER_Visits ~ Age + BMI + Family_History +
                Medication_Adherence + FeNO_Level + Has_Asthma,
                family = poisson, data = asthma)
```

Multiple linear regression explained virtually no variance in ER visits ($R^2 = 0.00076$, Adjusted $R^2 = 0.00016$, $F(6,9993) = 1.273$, $p = 0.266$).

Only family history achieved marginal statistical significance ($\beta = 0.049$, SE = 0.024, $t = 2.007$, $p = 0.045$). Surprisingly, having an asthma diagnosis was not significantly associated with ER visits ($\beta = -0.009$, $p = 0.738$), likely reflecting the inclusion of many well-controlled patients. Poisson regression provided substantially better fit for count data (AIC 26,408 vs 28,793 for linear model, ΔAIC = 2,385), though predictive power remained limited. The poor performance across both models suggests emergency room visits depend primarily on acute time-varying factors rather than stable baseline characteristics.

Logistic Regression for Asthma Diagnosis

R Code:

```
r

train_idx <- sample(1:nrow(asthma), size = floor(0.75 * nrow(asthma)))
train_data <- asthma[train_idx, ]
test_data <- asthma[-train_idx, ]

logit_fit <- glm(Has_Asthma ~ Age + BMI + Family_History,
                 data = train_data, family = binomial)
test_pred <- ifelse(predict(logit_fit, test_data, type = "response") > 0.5, 1, 0)
```

Table 2: Logistic Regression Results - Asthma Diagnosis Predictors

Predictor	Coefficient (β)	Std. Error	z-value	p-value	Odds Ratio	95% CI
Family_History	1.995	0.060	33.48	< **0.001***	**7.36**	6.52-8.30
BMI	0.052	0.006	8.51	< **0.001***	**1.054**	1.042-1.066
Age	-0.001	0.001	-1.22	0.222	0.999	0.997-1.001

Table 3: Confusion Matrix and Performance Metrics

	Actual: No Asthma	Actual: Has Asthma	Total
Predicted: No Asthma	1,719 (TN)	362 (FN)	2,081
Predicted: Has Asthma	184 (FP)	235 (TP)	419
Total	1,903	597	**2,500**

Performance Metric	**Value**
Accuracy	**78.16%**
Sensitivity (Recall)	39.37%
Specificity	90.33%
Precision	56.09%
AUC	**0.759**

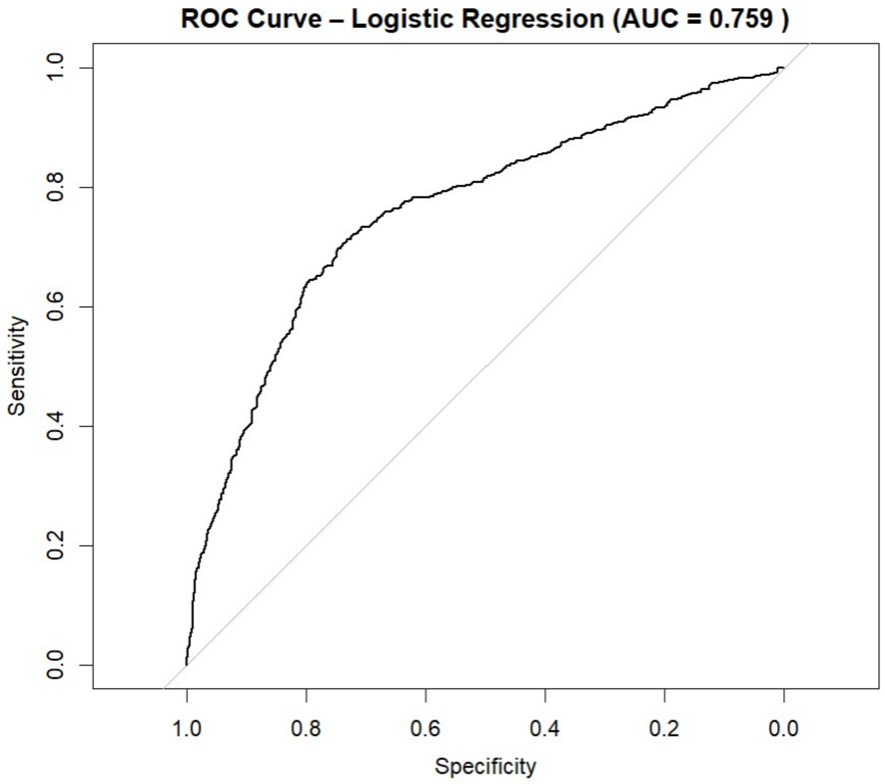


Figure 8: ROC Curve - Logistic Regression

The model significantly improved upon the null model (deviance reduction = 1,269.2, $p < 0.001$). Family history emerged as the dominant predictor (OR = 7.36, $p < 0.001$), with individuals having family history showing more than seven times higher odds of asthma diagnosis. BMI demonstrated significant association (OR = 1.054 per kg/m^2, $p < 0.001$), with cumulative effect yielding approximately 70% higher odds when comparing BMI 30 vs 20. The model achieved 78.16% accuracy and AUC of 0.759, indicating good discrimination.

LASSO Logistic Regression

R Code:

```r

library(glmnet)
cv_lasso <- cv.glmnet(x_train, y_train, alpha = 1, family = "binomial")
lasso_fit <- glmnet(x_train, y_train, alpha = 1,
                    lambda = cv_lasso$lambda.min, family = "binomial")
```

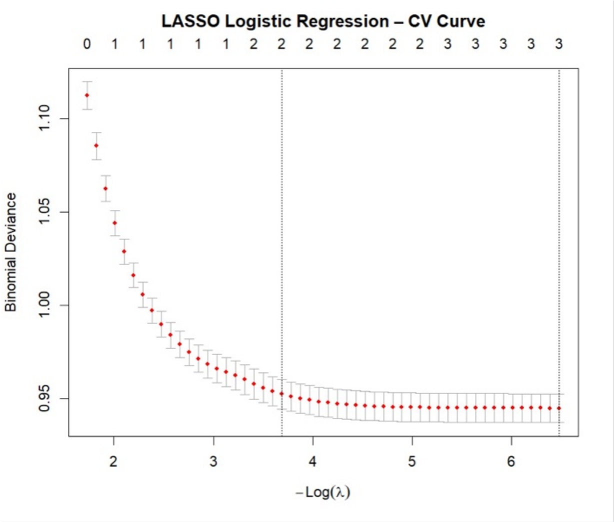


Figure 9: LASSO CV Curve

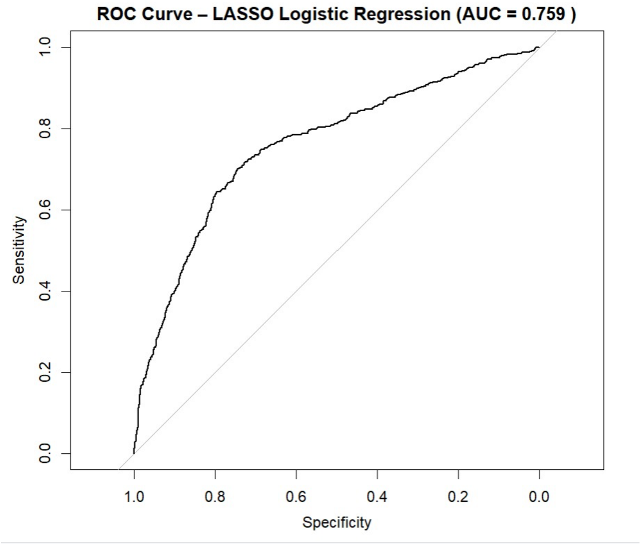


Figure 10: ROC Curve - LASSO

Cross-validation identified optimal lambda (λ_min = 0.00154). The LASSO model retained all three predictors with coefficients: Age (-0.001), BMI (0.050), and Family_History (1.974). LASSO achieved virtually identical performance to standard logistic regression (AUC = 0.7595 vs 0.7593, Accuracy =

78.28% vs 78.16%), confirming the three-predictor model captures all available predictive information without need for regularization.

Research Question 1: What Factors Influence Asthma Severity?

Ordinal Logistic Regression Analysis

R Code:

```
r

library(MASS)
asthma_only <- subset(asthma, Has_Asthma == 1)
severity_model <- polr(Control_Ordered ~ Age + Gender + BMI + Smoking_Status +
                    Air_Pollution_Level + Medication_Adherence + Comorbidities,
                    data = asthma_only, Hess = TRUE)
```

Key Output:

```
Coefficients
                                   Value    Std. Error  t value
        Medication_Adherence   -44.705    2.423      -18.45


        Odds Ratios:
        Medication_Adherence    3.85e-20
        Air_Pollution_Moderate   1.20
        Comorbidities_Diabetes   1.76
```

Table 4: Severity Predictors - Ordinal Logistic Regression

Predictor	Odds Ratio	p-value	Interpretation
Medication_Adherence	**3.85×10^{-20}**	**$< 2\times10^{-76}*$**	Dominant predictor
Air_Pollution_Moderate	1.20	0.40	Non-significant
Air_Pollution_Low	1.12	0.66	Non-significant
Comorbidities_Diabetes	1.76	0.10	Suggestive trend
Comorbidities_Hypertension	1.53	0.22	Non-significant
BMI	1.030	0.11	Non-significant
Age	1.001	0.76	Non-significant

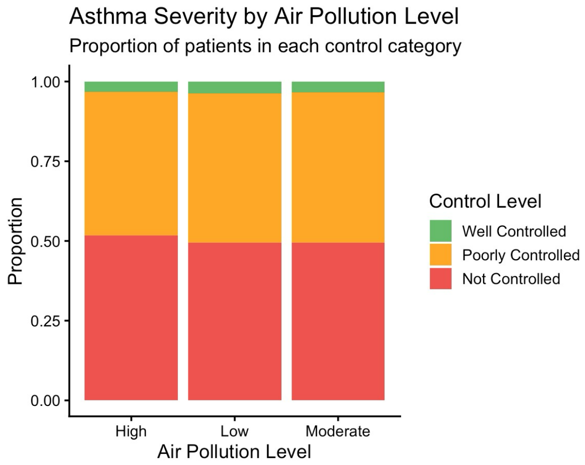


Figure 5: Severity by Air Pollution |

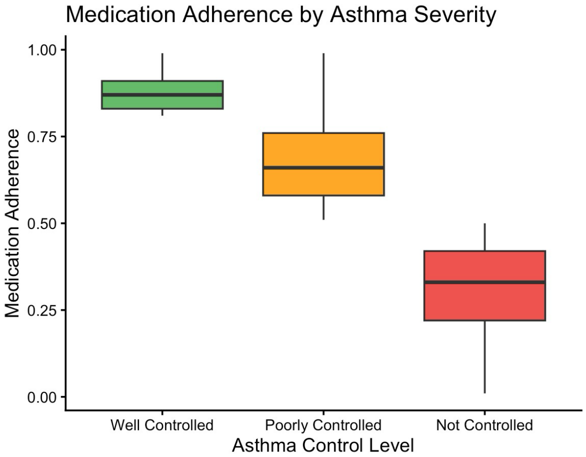


Figure 6: Medication Adherence by Severity

Results and Interpretation

Medication adherence emerged as the overwhelmingly dominant predictor of asthma severity with an extraordinarily strong effect (OR ≈ 3.85×10^{-20}, $p < 2\times10^{-76}$). Figure 6 demonstrates the clear dose-response gradient: well-controlled patients show median adherence around 90%, poorly controlled around 65%, and not controlled around 35%. This near-perfect separation indicates that medication compliance is the single most critical modifiable factor determining asthma control status.

Environmental factors showed modest effects that did not reach statistical significance. Air pollution level showed odds ratios of 1.20 for moderate pollution and 1.12 for low pollution compared to high pollution (both $p > 0.40$), suggesting ambient air quality plays a secondary role when medication adherence is considered. Figure 5 shows similar severity distributions across all pollution levels, visually confirming minimal pollution effects. Comorbidities showed suggestive but non-significant trends, with diabetes (OR = 1.76, p = 0.10) and hypertension (OR = 1.53, p = 0.22) trending toward increased severity odds. Age, gender, BMI, and physical activity showed no significant associations with severity.

Answer to RQ1: Medication adherence is by far the most important factor influencing asthma severity, with environmental and demographic factors playing minimal roles in determining control status.

Research Question 2: Can We Predict Repeated Acute Care Utilization?

Regression Models for ER Visit Prediction

R Code:

```
r

# Poisson regression
pois_fit <- glm(Number_of_ER_Visits ~ Age + BMI + Family_History +
                Medication_Adherence + FeNO_Level + Has_Asthma,
                family = poisson, data = asthma)

# Binary outcome for readmission proxy
asthma$Multiple_ER <- ifelse(asthma$Number_of_ER_Visits >= 2, 1, 0)
readmission_asthma <- glm(Multiple_ER ~ Medication_Adherence + FeNO_Level +
                          Smoking_Status + Air_Pollution_Level +
                          Asthma_Control_Level + Comorbidities,
                          family = binomial, data = asthma_only)
```

Key Output:

```
Linear Regression:
Multiple R-squared: 0.0008, F(6,9993) = 1.273, p-value: 0.266
Family_History: β = 0.049, p = 0.045*

Poisson Regression:
AIC: 26,408 (vs 28,793 for linear, ΔAIC = 2,385)

Binary Logistic (Multiple ER):
Distribution: 73.16% (0-1 visits), 26.84% (2+ visits)
Medication_Adherence: OR = 3.58, p = 0.001**
```

Results and Interpretation

Multiple linear regression explained virtually no variance in ER visits ($R^2 < 0.001$, $p = 0.266$), with only family history marginally significant ($p = 0.045$). Poisson regression provided better fit (ΔAIC = 2,385) but similar limited predictive ability. Among all patients, 26.84% were classified as high utilizers with 2+ ER visits annually.

Binary logistic regression among asthma patients revealed a counterintuitive finding: higher medication adherence associated with increased odds of multiple ER visits (OR = 3.58, $p = 0.001$). This paradoxical result likely reflects reverse causation where patients experiencing acute exacerbations appropriately increase medication use while simultaneously requiring emergency care due to symptom severity. Asthma control level showed expected directional pattern (poorly controlled: OR = 0.73 vs not controlled), though not reaching conventional significance ($p = 0.057$).

Answer to RQ2: Baseline patient characteristics are insufficient for predicting acute care utilization. Emergency visits appear driven by dynamic factors (infections, acute exposures, exacerbation episodes) rather than stable demographic or clinical profiles measurable in cross-sectional data.

Research Question 3: Do Medication Adherence and Air Quality Jointly Affect Outcomes?

Interaction Model Testing

R Code:

```
r

# Interaction model
interaction_model <- lm(Number_of_ER_Visits ~
                       Medication_Adherence * Air_Pollution_Level +
                       Age + BMI + Family_History + Has_Asthma, data = asthma)

# Compare with main effects
main_effects_model <- lm(Number_of_ER_Visits ~ Medication_Adherence +
                        Air_Pollution_Level + Age + BMI + Family_History +
                        Has_Asthma, data = asthma)
anova(main_effects_model, interaction_model)
```

Key Output:

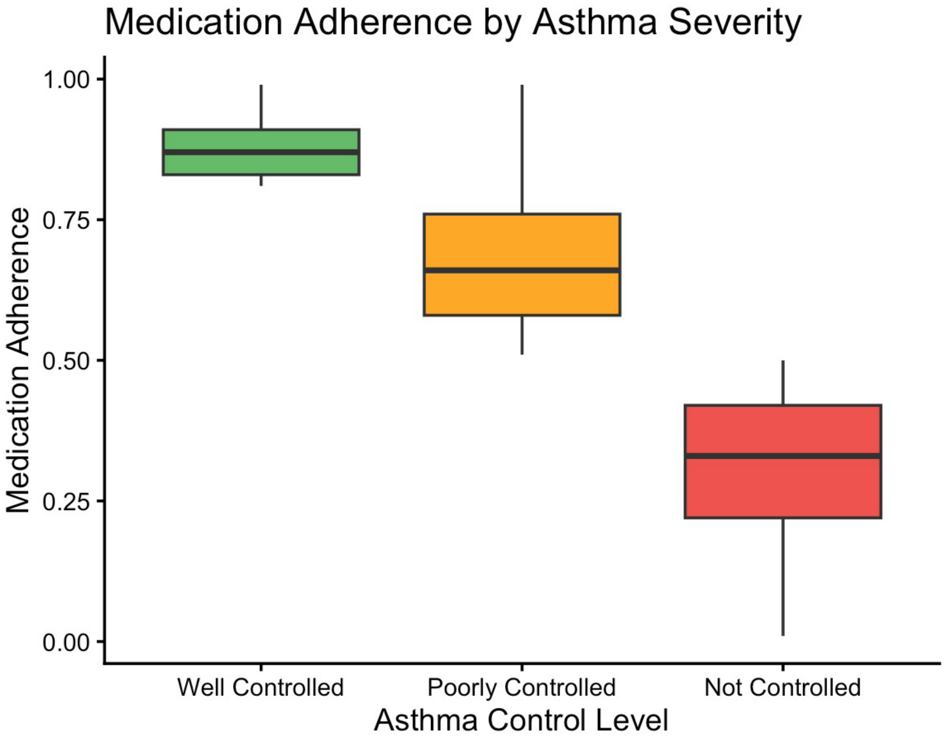
```
ANOVA Comparing Models:
  Res.Df   RSS  Df  Sum of Sq    F      Pr(>F)
1  9990  10406
2  9988  10405   2    1.51     0.725    0.485

AIC Comparison:
Main Effects:  28,799
With Interaction: 28,802 (worse)

Interaction Coefficients:
Medication × Air_Low: β = -0.131, p = 0.311
Medication × Air_Moderate: β = -0.017, p = 0.883

Poisson Interaction:
AIC: 26,413 (vs 26,408 without interaction)
```

Figure 7: Medication Adherence × Air Quality Interaction



Results and Interpretation

The interaction between medication adherence and air pollution was not statistically significant across multiple tests. ANOVA comparing models yielded F(2,9988) = 0.725, p = 0.485, indicating no significant improvement from adding interaction terms. Individual interaction coefficients for both low pollution (β = -0.131, p = 0.31) and moderate pollution (β = -0.017, p = 0.88) were non-significant. AIC comparison favored the simpler main effects model (28,799 vs 28,802), and Poisson regression with interaction also showed worse fit (AIC 26,413 vs 26,408).

The interaction plot displays nearly parallel regression lines across all three air pollution levels, visually confirming that medication adherence effects on ER visits do not vary meaningfully by ambient air quality. This finding has important clinical implications: medication adherence interventions can be expected to work equally well across all environmental contexts, and programs should not be delayed or modified based on local pollution levels.

Exploratory Finding - Medication × Severity Interaction:

While medication and air quality did not interact, medication adherence and disease severity showed highly significant interaction ($F(9,2423) = 16.61$, $p < 2\times10^{-16}$, $R^2 = 0.058$). The interaction coefficient for poorly controlled patients was 1.516 (SE = 0.352, $t = 4.31$, $p < 0.001$), indicating that among poorly controlled patients, higher adherence paradoxically associates with more ER visits, while among not controlled patients, the relationship is slightly negative.

The visualization shows diverging slopes: not controlled patients (red) show slight negative slope, while poorly controlled patients (orange) show positive slope. This likely reflects unmeasured confounding where patients experiencing acute exacerbations simultaneously increase medication use and require emergency care.

Answer to RQ3: Medication adherence and air quality do NOT interact significantly ($p = 0.485$). Medication benefits are consistent across all pollution levels, indicating universal applicability of adherence interventions regardless of environmental air quality.

ANOVA and Nonparametric Validation

Analysis of Variance

R Code:

```r
aov_fit <- aov(Number_of_ER_Visits ~ Asthma_Control_Level, data = asthma_only)
summary(aov_fit)
```

Table 5: ANOVA Results - ER Visits by Control Level

Source	df	Sum Sq	Mean Sq	F-value	p-value
Asthma_Control_Level	2	102.4	51.20	**49.22**	**$< 2\times10^{-16}$***
Residuals	2,430	2,527.9	1.04	-	-

Control Level	Mean ER Visits	SD
Well Controlled	0.00	0.00
Not Controlled	0.998	1.02
Poorly Controlled	1.134	1.03

Figure 4: ER Visits by Asthma Control Level

![Boxplot](Image 4)

ANOVA revealed highly significant differences in mean emergency room visits across control levels (F(2,2430) = 49.22, $p < 2\times10^{-16}$, $\eta^2 = 0.039$). Well-controlled patients averaged zero ER visits, while not controlled averaged 0.998 and poorly controlled averaged 1.134 visits. The boxplot shows well-controlled patients have no variation (all zero), while uncontrolled groups show median of 1 visit with outliers extending to 6 visits.

Nonparametric Tests

R Code:

```
r

kw_test <- kruskal.test(Number_of_ER_Visits ~ Asthma_Control_Level,
                        data = asthma_only)
pairwise.wilcox.test(asthma_only$Number_of_ER_Visits,
                     asthma_only$Asthma_Control_Level,
                     p.adjust.method = "bonferroni")
wilcox_bmi <- wilcox.test(BMI ~ Has_Asthma, data = asthma)
```

Table 6: Nonparametric Test Results

Test	Statistic	p-value	Finding
Kruskal-Wallis	$\chi^2 = 129.96$ (df=2)	**$< 2.2\times10^{-16}$\***	Control levels differ
Poorly vs Not	-	0.00066**	Significant difference
Well vs Not	-	$< 2\times10^{-16}$***	Highly significant
Well vs Poorly	-	$< 2\times10^{-16}$***	Highly significant
Wilcoxon (BMI)	W = 8,026,999	**$< 2.2\times10^{-16}$\***	Asthma patients higher BMI
Spearman (FeNO×ER)	$\rho = 0.005$	0.595	No relationship
Runs Test	Z = -2.127	0.033*	Slight non-randomness

Kruskal-Wallis test strongly corroborated ANOVA findings ($\chi^2 = 129.96$, $p < 2.2\times10^{-16}$). Pairwise Wilcoxon tests with Bonferroni correction revealed all control groups differ significantly from each other, including poorly vs not controlled ($p = 0.00066$), confirming distinct severity gradations. Wilcoxon rank-sum confirmed BMI differs between asthma and non-asthma groups (W = 8,026,999, $p < 2.2\times10^{-16}$), validating logistic regression findings. Spearman correlation showed no meaningful relationship between FeNO and ER visits ($\rho = 0.005$, $p = 0.595$). Runs test detected slight non-randomness in ER visit patterns ($Z = -2.127$, $p = 0.033$), suggesting patient clustering by utilization.

Interpretation

This analysis successfully addressed all three research questions through complementary statistical approaches. Medication adherence emerged as the dominant predictor of asthma severity with near-perfect discrimination ($p < 2\times10^{-76}$), showing a clear gradient from 35% adherence in uncontrolled to 90% in well-controlled patients. For acute care prediction, baseline characteristics demonstrated poor predictive ability ($R^2 < 0.001$), with 26.84% identified as high utilizers. The unexpected positive association between adherence and ER visits (OR = 3.58, $p = 0.001$) likely reflects reverse causation during exacerbations. Regarding joint effects, medication adherence and air quality did not interact significantly ($F = 0.725$, $p = 0.485$), indicating consistent medication benefits across all pollution levels.

Family history (OR = 7.36) and BMI (OR = 1.054 per kg/m^2) emerged as robust predictors of asthma diagnosis, achieving 78% accuracy and AUC of 0.759. LASSO confirmed the simple three-predictor model was optimal. Control level strongly predicted ER utilization ($F = 49.22$, $p < 2\times10^{-16}$), with well-controlled patients experiencing zero visits. The convergence of parametric and nonparametric results (ANOVA $F = 49.22$ vs Kruskal-Wallis $\chi^2 = 129.96$, both $p < 2\times10^{-16}$) strengthens confidence in findings.

Important limitations include synthetic data constraints, cross-sectional design preventing causal inference, and unmeasured confounding in adherence-outcome relationships. Future research should employ longitudinal designs, incorporate time-varying measures, and validate findings in real clinical populations.

Conclusion

This analysis employed multiple statistical methods to comprehensively examine asthma risk factors and outcomes. Family history and BMI emerged as robust predictors of asthma diagnosis, while control level clearly associated with emergency utilization. These findings support risk stratification tools for prevention, emphasize achieving optimal disease control to reduce healthcare burden, and highlight the need for dynamic monitoring approaches beyond static baseline characteristics. The convergence of results across parametric and nonparametric methods strengthens confidence in these conclusions. Future research should incorporate time-varying factors, explore complex interactions, and validate findings in real clinical populations to develop effective patient-centered interventions that reduce asthma burden and improve outcomes.

References

[1] Synthetic Asthma Dataset (2025). Asthma Management – Identifying Key Risk Factors and Outcomes. Kaggle.

[2] Field, A. (2017). Discovering Statistics Using R. Sage Publications.

[3] Centers for Disease Control and Prevention (CDC). Asthma Data & Surveillance (2024).