

Machine learning-based prediction of in-hospital mortality in patients with chronic respiratory disease exacerbations

DIGITAL HEALTH
Volume 11: 1–12
© The Author(s) 2025
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20552076251326703
journals.sagepub.com/home/dhj



Seung Yeob Ryu^{1,2,†}, Seon Min Lee^{2,†} , Young Jae Kim³  and Kwang Gi Kim^{4,5} 

Abstract

Objective: Exacerbation of chronic respiratory diseases leads to poor prognosis and a significant socioeconomic burden. To address this issue, an artificial intelligence model must assess patient prognosis early and classify patients into high- and low-risk groups. This study aimed to develop a model to predict in-hospital mortality in patients with chronic respiratory disease using demographic, clinical, and environmental factors, specifically air pollution exposure levels.

Methods: This study included 6272 patients diagnosed with chronic respiratory diseases comprising 39 risk factors. Air pollution indicators such as particulate matter (PM₁₀), fine particulate matter (PM_{2.5}), CO, NO₂, O₃, and SO₂ were used based on long-term and short-term exposure levels. Logistic regression, support vector machine, random forest, and extreme gradient boost were used to develop prediction models.

Results: The AUCs for the four models were 0.932, 0.935, 0.933, and 0.944. The key risk factors that significantly influenced predictions included blood urea nitrogen, red blood cell distribution width, respiratory rate, and age, which were positively correlated with mortality prediction. In contrast, albumin, lymphocyte count, diastolic blood pressure, and SpO₂ were negatively correlated with mortality prediction.

Conclusion: This study developed a prediction model for in-hospital mortality in patients with chronic respiratory disease and demonstrated a relatively high predictive performance. By incorporating environmental factors, such as air pollution exposure levels, the model with the best performance suggested that 365 days of exposure to air pollution was a key risk factor in mortality prediction.

Keywords

Chronic respiratory disease, mortality, prediction, machine learning, air pollution

Submission date: 25 July 2024; Acceptance date: 24 February 2025

Introduction

Asthma and chronic obstructive pulmonary disease (COPD) are the most common chronic respiratory diseases worldwide,¹ and the exacerbation of symptoms in chronic respiratory diseases is not only associated with poor prognosis but also with an increased socioeconomic burden due to treatment costs.² The mortality rate of patients with chronic respiratory diseases admitted to the intensive care unit is as high as 29%.³ In the United States, the direct costs associated with chronic respiratory diseases are estimated at \$32 billion annually, with indirect costs at \$20.4 billion annually.² To mitigate these risks, it is essential to perform early prognostic assessment of patients, classify them into high- and low-risk groups, and prevent unnecessary cost increases through timely treatment.³

¹Department of Biohealth & Medical Engineering, Gachon University, Seongnam-si, Gyeonggi-do, Korea

²Medical Devices R&D Center, Gachon University Gil Medical Center, Incheon, Korea

³Gachon Biomedical Convergence Institute, Gachon University Gil Medical Center, Incheon, Korea

⁴Department of Biomedical Engineering, College of Medicine, Gachon University Gil Medical Center, Incheon, Korea

⁵Department of Biomedical Engineering, College of IT Convergence, Gachon University, Seongnam-si, Gyeonggi-do, Korea

[†]Seung Yeob Ryu and Seon Min Lee contributed equally to this work. Seung Yeob Ryu and Seon Min Lee are the co-first (lead) authors.

Corresponding author:

Kwang Gi Kim, Department of Biomedical Engineering, College of Medicine, Gachon University Gil Medical Center, Incheon, Korea; Department of Biomedical Engineering, College of IT Convergence, Gachon University, Seongnam-si, Gyeonggi-do, Korea.

E-mail: kimkg@gachon.ac.kr



Numerous studies have been conducted to identify the risk factors for chronic respiratory diseases and predict symptom exacerbation based on these risk factors.

Risk factors for chronic respiratory diseases include demographic characteristics such as sex and age and clinical characteristics such as body measurements, medical history, vital signs, blood tests, and functional tests.^{4–7} Rezaee et al. identified significant risk factors for the acute exacerbation of chronic respiratory diseases, including patient age among demographic characteristics, respiration, pulse, smoking history, and medication use among clinical characteristics.⁵ Blood test results have been used as biomarkers for chronic respiratory diseases, including red blood cell, white blood cell, platelet, hemoglobin, hematocrit, and neutrophil and eosinophil counts.^{6,7}

These demographic and clinical characteristics have been used to predict exacerbation of chronic respiratory diseases. Goto et al. developed predictive models for the exacerbation of asthma and COPD in patients presenting to the emergency department using demographic and clinical characteristics, dividing patients into critical care and hospitalization groups.⁸ Gradient boosting machine (GBM) showed a concordance statistic (C-statistic) of 0.80 for critical care, and random forest (RF) showed a C-statistic of 0.83 for hospitalization.⁸ Shiroshita et al. developed and compared extreme gradient boost (XGB) and statistical models to predict in-hospital mortality of COPD patients, reporting AUCs of 0.71 and 0.69, respectively.⁹ Zein et al. developed models to predict hospitalizations due to symptom exacerbation in asthma and COPD patients, reporting AUCs of 0.81, 0.79, and 0.85 for logistic regression (LR), RF, and LightGBM, respectively.¹⁰

The significance of previous studies lies in the proposal of machine learning models that are superior to simple statistical analyses.¹¹ Predictive models using machine learning have the advantage of obtaining stable predictive results compared to statistical models by considering non-linear interactions between independent variables from large datasets.^{8–10} However, the models in previous studies have low performance for clinical applications and lack universality, as they arbitrarily divide patient groups to develop individual predictive models. Additionally, they are limited as they predict exacerbations using only demographic and clinical characteristics without considering patients' environmental characteristics.

The environmental characteristics of patients include economic level, presence of a caregiver, residential area, and degree of exposure to air pollution, with the latter showing a high correlation with respiratory diseases.^{4,12,13} Furthermore, fine particulate matter, an indicator of air pollution, affects acute exacerbation and 30-day readmission regardless of the disease.¹⁴ However, no studies have evaluated the prognosis of the most vulnerable respiratory diseases or predicted symptom exacerbations using these air pollution concentrations. Therefore, a predictive model

applicable to patients with chronic respiratory diseases that includes statistically significant risk factors and environmental risk factors related to patients, such as the degree of exposure to air pollution, is needed.

In this study, we aimed to predict in-hospital mortality due to symptom exacerbation of chronic respiratory diseases using patients' demographic and clinical characteristics as well as environmental characteristics, such as the degree of exposure to air pollution. By utilizing machine learning, we propose a model that aids therapeutic decision-making in clinical settings by early assessment of the prognosis of patients diagnosed with and hospitalized for chronic respiratory diseases and categorizing patients into high-risk and low-risk groups based on in-hospital mortality predictions.

Methods

Development environment

In this study, an Intel® Core™ i9-10900 (Intel, Santa Clara, CA, USA) system with 32 GB RAM was used, and experiments were conducted using Python (version 3.7.0, Python Software Foundation, Wilmington, DE, USA) on a 64-bit CPU. Model training was performed using a framework based on Scikit-learn (version 1.0.2), Imblearn (version 0.13.0) and statistical analysis was performed using MedCalc (Version 19.6.1, MedCalc Software, Ostend, Belgium).

Data collection

This study was designed as a retrospective chart review study and was conducted at Gachon University Gil Medical Center in Incheon, South Korea. We collected electronic health record data from 6272 patients diagnosed with chronic respiratory diseases as their primary diagnosis, who were admitted to Gachon University Gil Medical Center and resided in Incheon from 1 January 2019 to 31 December 2023. The patients included 628 in-hospital mortality and 5644 control who were discharged normally after treatment during the same period. The in-hospital mortality group data was retrospectively collected for patients who died within 30 days of hospitalization. The entire dataset was randomly split into training ($n=5017$) and test ($n=1255$) datasets in a ratio of 8:2. The primary diagnostic codes of the patients were based on the International Classification of Diseases (ICD 10), as shown in Supplementary Materials 1. All data related to patient information were collected from the Clinical Research Data Warehouse (CRDW) after obtaining approval from the Institutional Review Board (GBIRB2024-167) at Gachon University Gil Medical Center. Information on air pollution in the patients' residential areas was collected through the Environmental Information Disclosure System (<https://air.>

incheon.go.kr/) using measurement records from regional urban air quality monitoring stations.

Patient data comprised 30 clinical factors and nine non-clinical factors based on the risk factors that affect chronic respiratory diseases (Supplementary Materials 2). Clinical factors include demographic characteristics, severity, acute clinical stability, and physical functional status, which can be identified upon patient admission.⁴ In this study, clinical factors included age, gender, vital signs (blood pressure,¹⁵ body temperature, pulse,¹⁶ respiration,¹⁶ oxygen saturation,¹⁶ and laboratory values (Hb,¹⁷ Hct,¹⁷ PLT, neutrophils,¹⁸ lymphocytes,¹⁹ basophils, monocytes, eosinophils,¹⁸ MCV,¹⁷ MCH,¹⁷ MCHC,¹⁷ MPV,²⁰ PDW,²¹ red blood cell distribution width (RDW),^{22,23} sodium,²⁴ potassium, calcium,²⁵ total bilirubin, blood urea nitrogen (BUN),^{26,27} creatinine, albumin,^{28,29} and hsCRP.¹⁷ Non-clinical factors include psychological, cognitive, and social functioning; cultural, ethnic, and socioeconomic beliefs and behaviors; and health-related quality of life, which are more indirectly related to health status or functional level than clinical factors.⁴ Non-clinical factors included the patient's residential area, marital status, smoking status, and concentration of air pollutants in the residential area, specifically particulate matter (PM10), fine particulate matter (PM2.5), CO, NO₂, O₃, and SO₂. In this study, the 34 factors mentioned previously were used as independent variables, and the patient's in-hospital mortality was used as the dependent variable.

Data preprocessing

All patients underwent repeated vital sign measurements and hematological tests during hospitalization, resulting in a large amount of data on vital signs and laboratory values. Therefore, the mean value of each variable was calculated. For air pollution indicators, PM10, PM2.5, CO, NO₂, O₃, and SO₂ were used, and the patients' residential areas were used at the district level. Each indicator was pre-processed into long-term and short-term exposure levels by residential area. The long-term exposure level was calculated as the average value over 365 days based on the patient's admission date. Short-term exposure levels were calculated as the average values over 7 and 3 days based on the patient's admission date.

Data were separated into categorical and numerical variables (Supplementary Materials 2). Categorical variables were mapped to integers corresponding to each category and then converted to the numerical value of that category. Subsequently, categorical and numerical variables were normalized to a range of 0–1 for application to the machine learning models. The training dataset was under-sampled using the RandomUnderSampler algorithm from the Imblearn library to prevent model overfitting due to data imbalance (n = 1004), ensuring that the mortality and control groups were trained at an equal ratio (Figure 1).

Model training

In this study, the machine learning classification models used were LR, support vector machine (SVM), RF, and XGB. LR is a basic and effective machine learning method suitable for binary classification and modeling probabilities based on a sigmoid function.³⁰ SVM, developed by Vapnik and Chervonenkis in 1963, is a classification model that can transform nonlinear inputs into a linear state, depending on the kernel function used.^{31,32} RF, an ensemble model developed by Breiman in 2001, prevents overfitting by expanding processing based on the amount of information while maintaining statistical efficiency.^{33,34} XGB is an enhanced version of the traditional Gradient Tree Boosting algorithm that incorporates techniques to prevent overfitting and has the advantage of fast classification through parallel processing.³⁵ As shown in Figure 1, the training parameters of the four models were optimized using a grid-search technique for hyperparameter tuning. The optimized parameters are listed in Supplementary Materials 3.

Statistical analysis

Continuous data among the independent variables were analyzed using an independent sample t-test to test the statistical significance between the control and expired groups. Additionally, all independent variables were quantitatively evaluated for their impact on mortality prediction using permutation importance.³⁶ The permutation importance was calculated using an algorithm provided by Scikit-learn, and the importance (i_j) of each independent variable (j) was calculated using the model's performance metric (s), specifically the AUC, as shown in the following formula:

$$i_j = s - \frac{1}{K} \sum_{k=1}^K s_{k,j}$$

Subsequently, to analyze the correlation between each independent variable and the prediction results of the four models, Shapley additive explanations (SHAP) values are presented.^{37,38} The SHAP value method has emerged to address the black-box problem inherent in traditional artificial intelligence models. It has the advantage of indicating the influence and directionality of each variable.³⁸

Results

A total of 2259 participants were used to train and validate the models. Fivefold cross-validation was performed on the training dataset (n = 1004) using the grid-search technique, and the models trained with the optimal parameters were evaluated using the test dataset (n = 1255). The statistical characteristics of the clinical and non-clinical factors

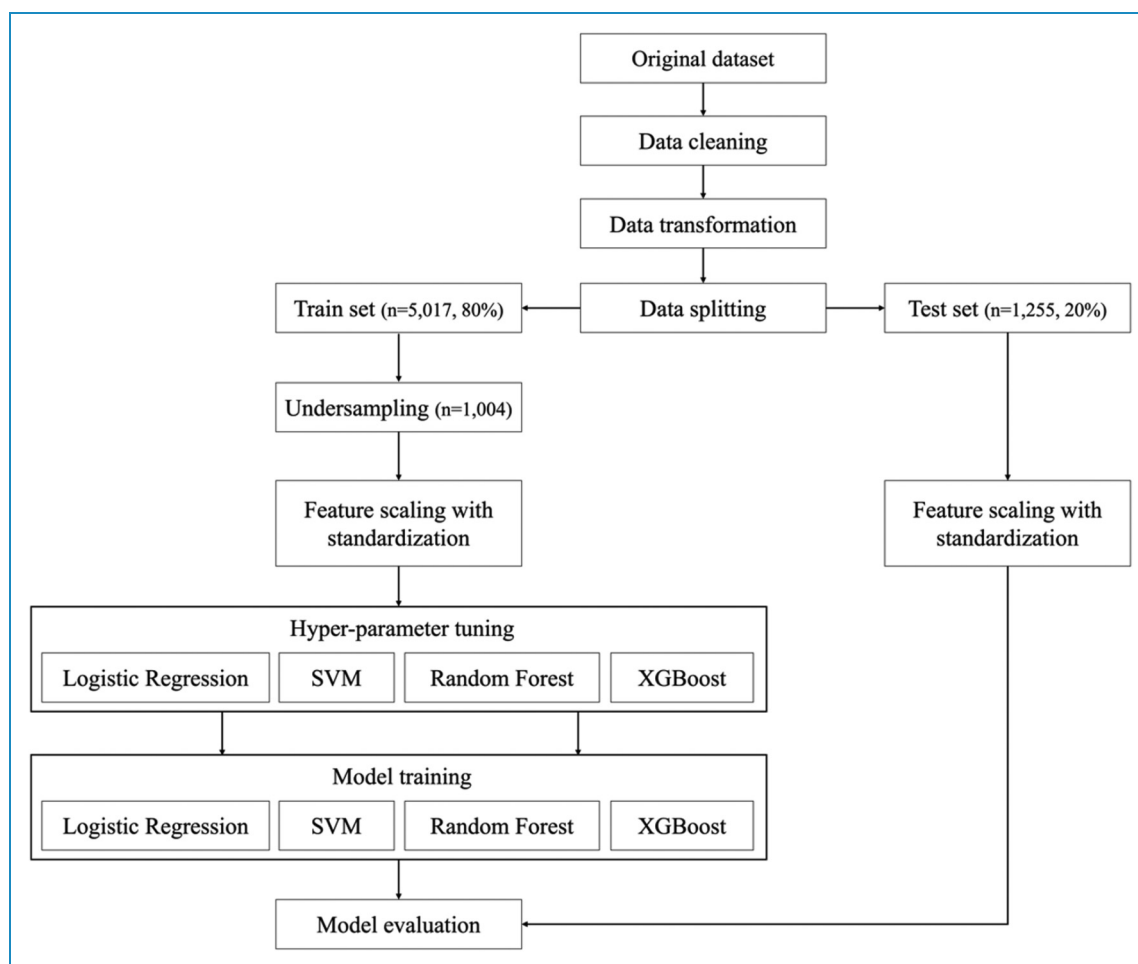


Figure 1. Flow chart of data preprocessing and modeling.

collected from the subjects are presented in Tables 1 and 2, respectively. The average ages of the control and expired groups were 75 and 83 years, respectively, with males constituting 52.4% and 64.0% of each group. The most common residential areas in both groups were the same. The majority were married (86.1% and 87.3% in the control and expired groups, respectively), and nonsmokers accounted for 80.4% and 85.2%, respectively. There were statistically significant differences between the control and expired groups in most clinical factors, except for body temperature and MCH, and in non-clinical factors, precisely the 365-day average values of CO, NO₂, and O₃ and the 7 days average value of NO₂ ($p < 0.05$).

The performance of each model was validated using a separately constructed test dataset. The performance of the LR, SVM, RF, and XGB models was evaluated in terms of sensitivity, specificity, accuracy, and AUC. Sensitivity, specificity, accuracy, and AUC were calculated using a confusion matrix, and the AUC was derived from the receiver operating characteristic curve based on the confusion matrix, presenting the area under the curve.³⁹ The

results are presented in Table 3 and Figure 2. Comparing the performance of each model, XGB demonstrated a higher AUC than the other models, with a statistically significant difference compared to LR ($p = 0.044$).

All independent variables were quantitatively evaluated for their impact on mortality prediction using permutation importance.³⁶ Figure 3 shows the top 10 variables by average importance calculated for the LR, SVM, RF, and XGB models. Figure 4 displays the important variables of all the models and their respective SHAP values, illustrating the correlation between each variable and the prediction outcomes. Based on the analysis of the average importance and model-specific SHAP values, BUN, albumin, RDW, respiratory rate, lymphocyte count, age, diastolic blood pressure (DBP), and SpO₂ emerged as common risk factors. BUN, RDW, respiratory rate, and age positively correlated with mortality prediction, whereas albumin levels, lymphocyte counts, DBP, and SpO₂ negatively correlated with mortality prediction.

Table 1. Clinical characteristics of patients in chronic respiratory disease.

Variable	In-hospital mortality						p-value
	No (n = 1631)			Yes (n = 628)			
	Mean or n (%)	Std.	Min-max	Mean or n (%)	Std.	Min-max	
Age (yr)	71.8	15.2	7–101	81.3	9.6	40–105	< 0.001
Gender (Male)	854 (52.4%)	-	-	402 (64.0%)	-	-	
SBP (mmHg)	126.9	12.5	87.8–178.0	119.3	16.3	10.0–161.0	< 0.001
DBP (mmHg)	74.9	8.2	39.3–112.0	67.6	10.5	8.0–110.0	< 0.001
Temperature (°C)	36.9	0.3	34.7–38.6	36.9	0.4	33.8–38.6	0.016
Pulse rate (per min)	76.7	10.8	6.0–143.3	88.6	16.6	0.0–152.0	< 0.001
Respiratory rate (per min)	19.9	2.2	10.0–94.0	20.9	4.8	11.3–104.0	< 0.001
SpO2 (%)	96.5	1.4	86.0–100.0	95.5	2.7	75.0–100.0	< 0.001
Hb (g/dL)	11.8	1.9	6.0–18.3	10.0	1.9	5.0–20.7	< 0.001
Hct (%)	35.4	5.4	17.6–54.6	30.7	5.7	9.1–61.8	< 0.001
PLT (×10 ³ /L)	227.6	81.5	22.0–818.0	191.7	100.8	16.0–577.9	< 0.001
Neutrophils (/μL)	66.1	11.5	4.6–96.3	81.1	11.3	4.0–98.4	< 0.001
Lymphocyte (%)	22.8	9.9	2.0–65.0	10.1	7.3	0.4–65.0	< 0.001
Basophils (%)	0.5	0.4	0.0–13.2	0.3	0.2	0.0–3.4	< 0.001
Monocyte (%)	7.9	2.4	0.5–19.441	6.3	3.3	0.4–27.1	< 0.001
Eosinophils (%)	2.6	2.5	0.0–28.3	1.6	2.6	0.0–31.5	< 0.001
Absolute neutrophils Counts	5525.8	3477.6	71.0–95167.0	10846.1	6690.0	75.0–51612.2	< 0.001
Absolute lymphocyte Counts	1611.0	774.2	123.7–16694.5	972.6	601.5	80.0–5727.0	< 0.001
MCV (fL)	91.5	5.3	62.6–116.7	93.1	6.1	67.9–113.4	< 0.001
MCH (pg)	30.5	2.1	17.0–38.7	30.5	2.3	20.2–37.8	0.601
MCHC (g/dL)	33.3	1.1	27.2–37.5	32.8	1.4	28.2–37.5	< 0.001
MPV (fL)	10.0	0.8	7.8–14.0	10.7	1.1	7.5–14.4	< 0.001
PDW (fL)	11.2	3.2	7.0–55.3	12.7	4.1	7.7–55.9	< 0.001
RDW (%)	13.4	1.7	11.0–23.1	15.6	2.4	11.0–29.1	< 0.001
Sodium (mmol/L)	139.3	3.2	119.5–153.0	137.7	5.8	107.0–165.5	< 0.001

(continued)

Table 1. Continued.

Variable	In-hospital mortality						p-value
	No (n = 1631)			Yes (n = 628)			
	Mean or n (%)	Std.	Min-max	Mean or n (%)	Std.	Min-max	
Potassium (mmol/L)	4.0	0.4	2.5–6.2	4.2	0.7	2.0–7.6	< 0.001
Calcium (mg/dL)	8.4	0.6	6.0–10.8	7.9	0.8	4.7–11.2	< 0.001
Total bilirubin (mg/dL)	0.7	0.6	0.2–11.4	1.6	3.1	0.2–27.9	< 0.001
BUN (mg/dL)	17.6	9.9	4.4–112.2	37.2	23.5	4.3–158.1	< 0.001
Creatinine (mg/dL)	0.9	1.0	0.2–13.7	1.5	1.3	0.2–12.8	< 0.001
Albumin (g/dL)	3.7	0.5	1.9–5.2	2.9	0.5	1.6–4.7	< 0.001
hsCRP (mg/dL)	3.3	3.8	0.0–22.0	10.0	7.4	0.1–40.5	< 0.001

SBP: systolic blood pressure; DBP: diastolic blood pressure; SpO₂: saturation pulse oxygen; Hb: hemoglobin; Hct: hematocrit, PLT: platelet count; MCV: mean corpuscular volume; MCH: mean corpuscular hemoglobin; MCHC: mean corpuscular hemoglobin concentration; MPV: mean platelet volume; PDW: platelet distribution width; RDW: red blood cell distribution width; BUN: blood urea nitrogen; hsCRP: high-sensitivity C-reactive protein.

Discussion

This study proposes a machine learning-based model to predict in-hospital mortality due to the exacerbation of symptoms in patients with chronic respiratory disease using demographic, clinical, and environmental characteristics. Predictions were made using the LR, SVM, RF, and XGB models, and the predictive performance was evaluated using sensitivity, specificity, accuracy, and AUC. The XGB model demonstrated superior performance compared with the other models.

The key variables influencing mortality prediction included BUN, RDW, respiratory rate, age, albumin level, lymphocyte count, DBP, and SpO₂. BUN, RDW, respiratory rate, and age were positively correlated with mortality prediction (Figure 4), consistent with previous studies.^{9,22,23,26,27} High BUN levels are strongly associated with mortality from various diseases and have been identified as a major risk factor for exacerbation and in-hospital mortality in COPD.²⁶ The critical BUN threshold directly contributing to COPD patient mortality is 7.30–7.63 mmol/L (approximately 131.40–137.34 mg/dL),^{26,27} and the maximum value of the expired group in this study exceeded this threshold. High RDW has been suggested as a risk factor for respiratory, cardiovascular, and hematological conditions.^{22,23} Previous studies have shown that an increase in the RDW is highly correlated with acute mortality in patients with COPD in the stable phase.²² The normal range for RDW is 11.8–14.3%²³; the control group in this study had an average RDW of 13.4%, whereas the

expired group had an average RDW of 15.6%. Typical clinical symptoms of COPD include tachypnea, tachycardia, and low oxygen saturation,¹⁶ and the expired group in this study showed statistically significant differences in respiratory rate, pulse rate, and SpO₂ compared to the control group ($p < 0.001$). These factors contributed significantly to the performance of the predictive model, as shown in Figures 3 and 4.

Variables negatively correlated with mortality prediction included SpO₂, serum albumin, lymphocytes, and DBP, consistent with previous studies.^{15,19,28,29,40,41} Serum albumin is a negative acute-phase reactant and a representative clinical indicator of malnutrition.²⁸ Malnutrition is a common comorbidity in patients with COPD that leads to lower serum albumin levels.^{28,29} The expired group had a significantly lower average serum albumin level than the control group ($p < 0.001$). Hematological indicators such as lymphocytes, monocytes, and eosinophils are closely related to lung function.^{19,40} Among these, a decrease in lymphocyte count is highly correlated with the exacerbation of COPD symptoms.¹⁹ The average lymphocyte counts in the control and expired groups were 22.8% and 10.1%, respectively, with a statistically significant difference ($p < 0.001$). Blood pressure was measured as SBP and DBP, with normal ranges of 80–140 mmHg and 60–90 mmHg, respectively.^{15,41} Both blood pressures showed a U-shaped relationship with mortality from all diseases, indicating that mortality risk increases when blood pressure deviates from the normal range.¹⁵ Previous studies have shown that low DBP is closely associated with in-hospital mortality in patients.⁴¹ In this study,

Table 2. Non-clinical characteristics of patients in chronic respiratory disease.

Variable	In-hospital mortality						p-value
	No (n = 1631)			Yes (n = 628)			
	Mean or n (%)	Std.	Min-max	Mean or n (%)	Std.	Min-max	
Residence (Khu)	699 (42.8%)	-	-	302 (48.1%)	-	-	
Marital status (yes)	1405 (86.1%)	-	-	548 (87.3%)	-	-	
Smoking (yes)	1311 (80.4%)	-	-	535 (85.2%)	-	-	
Average for 365 days							
CO (ppm)	0.514	0.060	0.256-0.670	0.521	0.06	0.266-0.668	0.009
NO ₂ (ppm)	0.024	0.004	0.008-0.032	0.024	0.004	0.008-0.032	0.001
O ₃ (ppm)	0.029	0.003	0.022-0.046	0.028	0.003	0.022-0.044	0.001
SO ₂ (ppm)	0.004	0.001	0.002-0.006	0.004	0.001	0.002-0.006	0.537
PM10 (µg/m ³)	38.2	3.9	30.2-50.0	37.9	3.7	31.1-48.0	0.078
PM2.5 (µg/m ³)	21.1	2.2	16.3-28.6	20.9	2.3	16.4-28.4	0.083
Average for 7 days							
CO (ppm)	0.511	0.140	0.186-1.224	0.520	0.142	0.214-1.124	0.169
NO ₂ (ppm)	0.023	0.009	0.004-0.060	0.024	0.009	0.004-0.056	0.016
O ₃ (ppm)	0.029	0.011	0.006-0.066	0.029	0.011	0.008-0.057	0.118
SO ₂ (ppm)	0.004	0.001	0.002-0.010	0.004	0.001	0.002-0.008	0.892
PM10 (µg/m ³)	38.4	18.3	7.0-139.2	38.7	18.9	10.5-148.6	0.716
PM2.5 (µg/m ³)	21.3	10.4	3.0-95.0	20.7	9.0	5.0-77.7	0.181
Average for 3 days							
CO (ppm)	0.514	0.159	0.200-1.389	0.519	0.157	0.200-1.367	0.508
NO ₂ (ppm)	0.023	0.011	0.003-0.063	0.024	0.011	0.004-0.061	0.042
O ₃ (ppm)	0.029	0.012	0.003-0.079	0.029	0.011	0.006-0.068	0.122
SO ₂ (ppm)	0.004	0.001	0.002-0.010	0.004	0.001	0.002-0.009	0.945
PM10 (µg/m ³)	38.4	21.5	6.0-258.0	38.6	23.7	9.2-264.5	0.822
PM2.5 (µg/m ³)	21.5	12.6	3.0-113.0	20.7	11.4	3.5-89.9	0.125

SBP: systolic blood pressure; DBP: diastolic blood pressure; SpO₂: saturation pulse oxygen; Hb: hemoglobin; Hct: hematocrit; PLT: platelet count; MCV: mean corpuscular volume; MCH: mean corpuscular hemoglobin; MCHC: mean corpuscular hemoglobin concentration; MPV: mean platelet volume; PDW: platelet distribution width; RDW: red blood cell distribution width; BUN: blood urea nitrogen; hsCRP: high-sensitivity C-reactive protein.

Table 3. Result of performance of all models.

	Sensitivity (95% CI)	Specificity (95% CI)	Accuracy (95% CI)	AUC (95% CI)
LR	0.817 (0.739–0.881)	0.864 (0.843–0.884)	0.860 (0.839–0.878)	0.932 (0.917–0.945)
SVM	0.849 (0.774–0.907)	0.865 (0.844–0.885)	0.864 (0.844–0.882)	0.935 (0.920–0.948)
RF	0.857 (0.784–0.913)	0.861 (0.839–0.881)	0.861 (0.840–0.879)	0.933 (0.918–0.946)
XGB	0.841 (0.766–0.900)	0.875 (0.854–0.894)	0.872 (0.852–0.890)	0.944 (0.930–0.956)

LR: logistic regression; SVM: support vector machine; RF: random forest; XGB: extreme gradient boost.

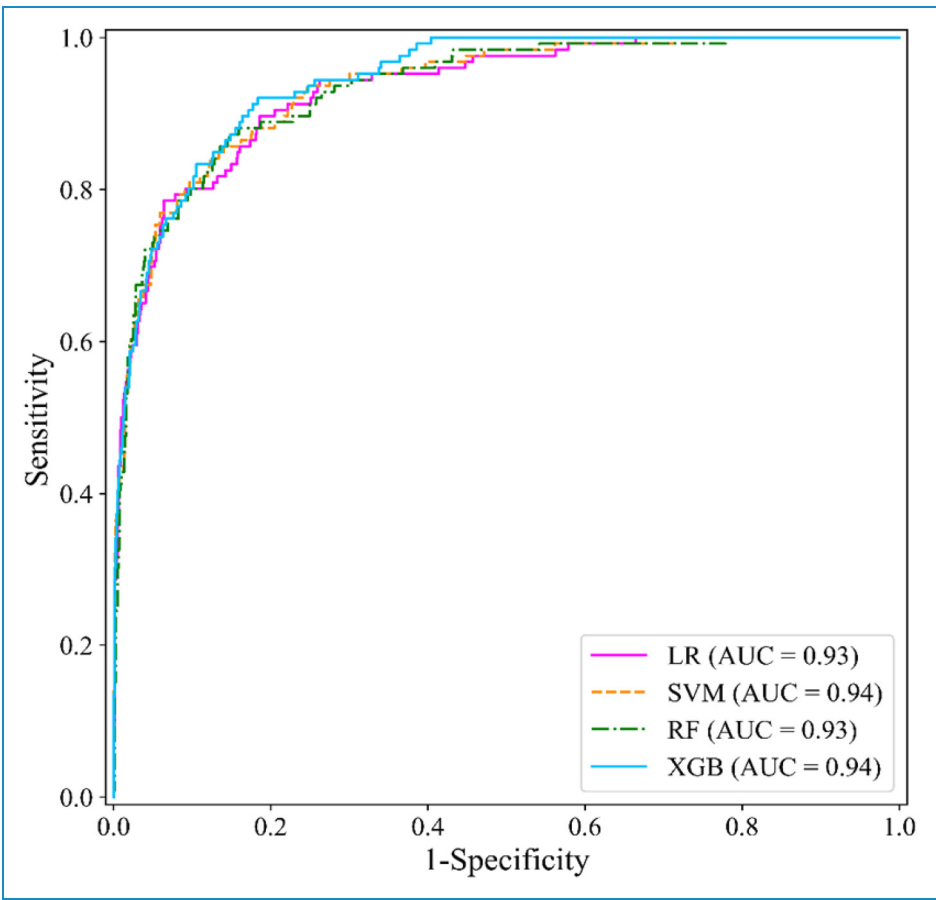


Figure 2. Comparison of ROC curves about prediction models.

the average DBP was 74.9 mmHg, while that of the expired group was 67.6 mmHg, which falls within the normal range; however, the difference between the two groups was statistically significant ($p < 0.001$).

As with the variables mentioned earlier, demographic and clinical variables have been used in previous studies to develop models predicting the prognosis of chronic respiratory diseases.^{8–10} Goto et al. used approximately 45 demographic and clinical variables, focusing on medical history, comorbidities, and chief complaints.⁸

Patients were divided into critical care and hospitalization groups for prediction, with GBM showing a C-statistic of 0.80 and critical care and RF showing a C-statistic of 0.83 for hospitalization.⁸ Shiroshita et al. compared the BAP-65 and CURB-65 models with machine learning models and applied variables such as age, consciousness level, pulse rate, respiratory rate, SBP, DBP, and BUN required by BAP-65 and CURB-65 to machine learning.⁹ The XGB model performed best with an AUC of 0.71.⁹ Zein et al. used 56 demographic and clinical variables,

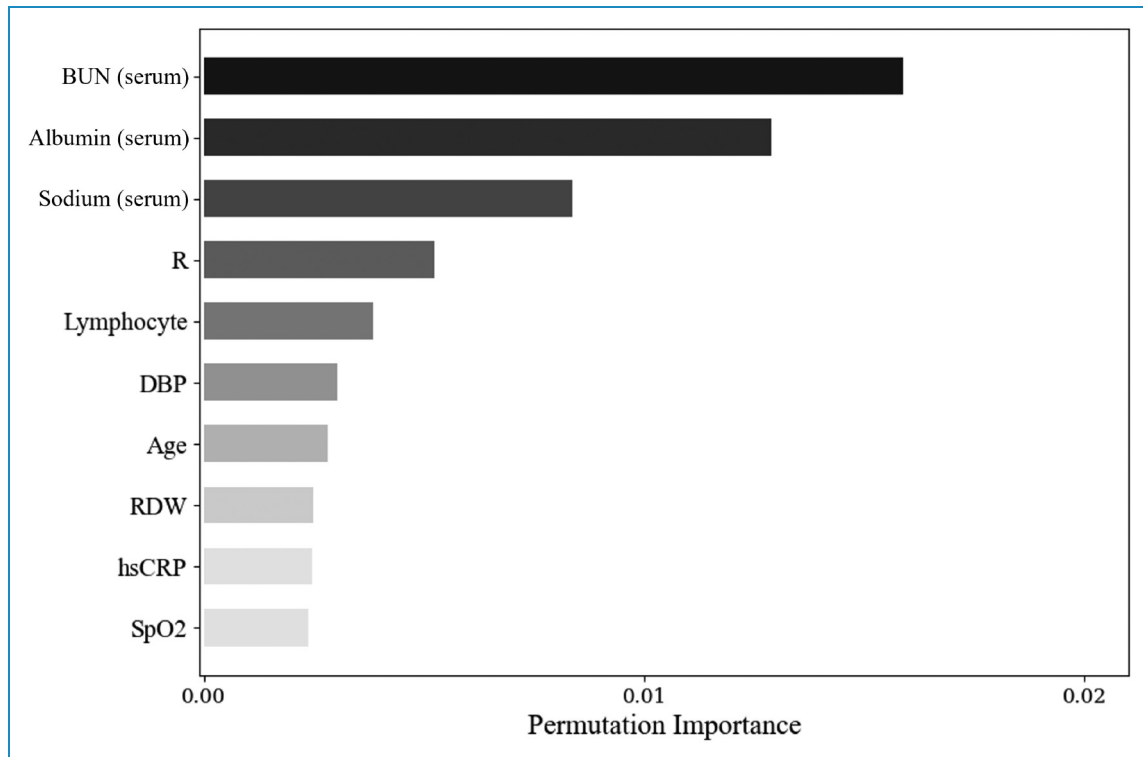


Figure 3. Top 10 prediction factors based on average permutation importance across models, R: respiratory rate per min.

including medication history and drug type.¹⁰ Of the LR, RF, and LightGBM models, LightGBM showed the best performance, with an AUC of 0.85.¹⁰ Unlike previous studies, the clinical variables in this study were limited to CBC and general chemistry tests (serum). The CBC is the primary test most commonly requested by clinicians in all clinical settings, offering simplicity, low cost, and feasibility for all medical facilities.⁴² To perform a CBC test, blood is drawn from the patient and stored in an SST tube, which allows serum tests to be conducted.⁴² Thus, most inpatients undergo CBC and serum tests; the results are recorded in the EMR. The clinical variables in this study were extracted from EMR-based CBC and serum test results. In contrast, demographic variables were collected from EMR-based patient information during initial nursing assessments upon admission.

Furthermore, this study includes environmental characteristics, specifically air pollution exposure levels, as variables in the prediction model. Previous studies have shown that long-term exposure to air pollution is associated with various systemic diseases, such as cardiovascular, respiratory, and neurological diseases,⁴³ with a particularly high correlation with the onset of bronchial disease.^{12,13} Short-term exposure to air pollution has been identified as a risk factor affecting 30-day readmission rates, regardless of the disease type.¹⁴ Long-term exposure levels in previous studies were defined as annual averages based on the admission date.⁴³ In contrast, short-term exposure levels were

defined as the 7-day average based on the admission date.¹⁴ This study used the annual, 7-day, and 3-day averages based on the admission date to examine the correlation with the prediction results for each period. Long-term exposure was a more significant factor than short-term exposure in predicting in-hospital mortality in patients with chronic respiratory disease (Figure 4). Among the air pollution indicators, long-term exposure to CO, NO₂, and O₃ showed statistically significant differences between the control and expired groups ($p < 0.01$), with CO and NO₂ positively correlated with mortality prediction (Figure 4). These results align with those of previous studies, emphasizing the importance of air pollution exposure levels.^{12–14,43} In conclusion, this study confirmed that the environmental characteristics of patients are important for early assessment of chronic respiratory disease prognosis.

This study has several strengths. Developing an in-hospital mortality prediction model that reflects the environmental characteristics of patients with chronic respiratory diseases is important. It utilizes all air pollution indicators, highlighting the risks of long-term exposure to CO and NO₂, and distinguishes itself from previous studies that did not use environmental characteristics.^{8–10} It also has the advantage of generalizability, covering a broader set of diseases than previous studies and performing well in predicting mortality in several chronic respiratory diseases, including COPD, asthma, and bronchiectasis. Second, the proposed model is applicable

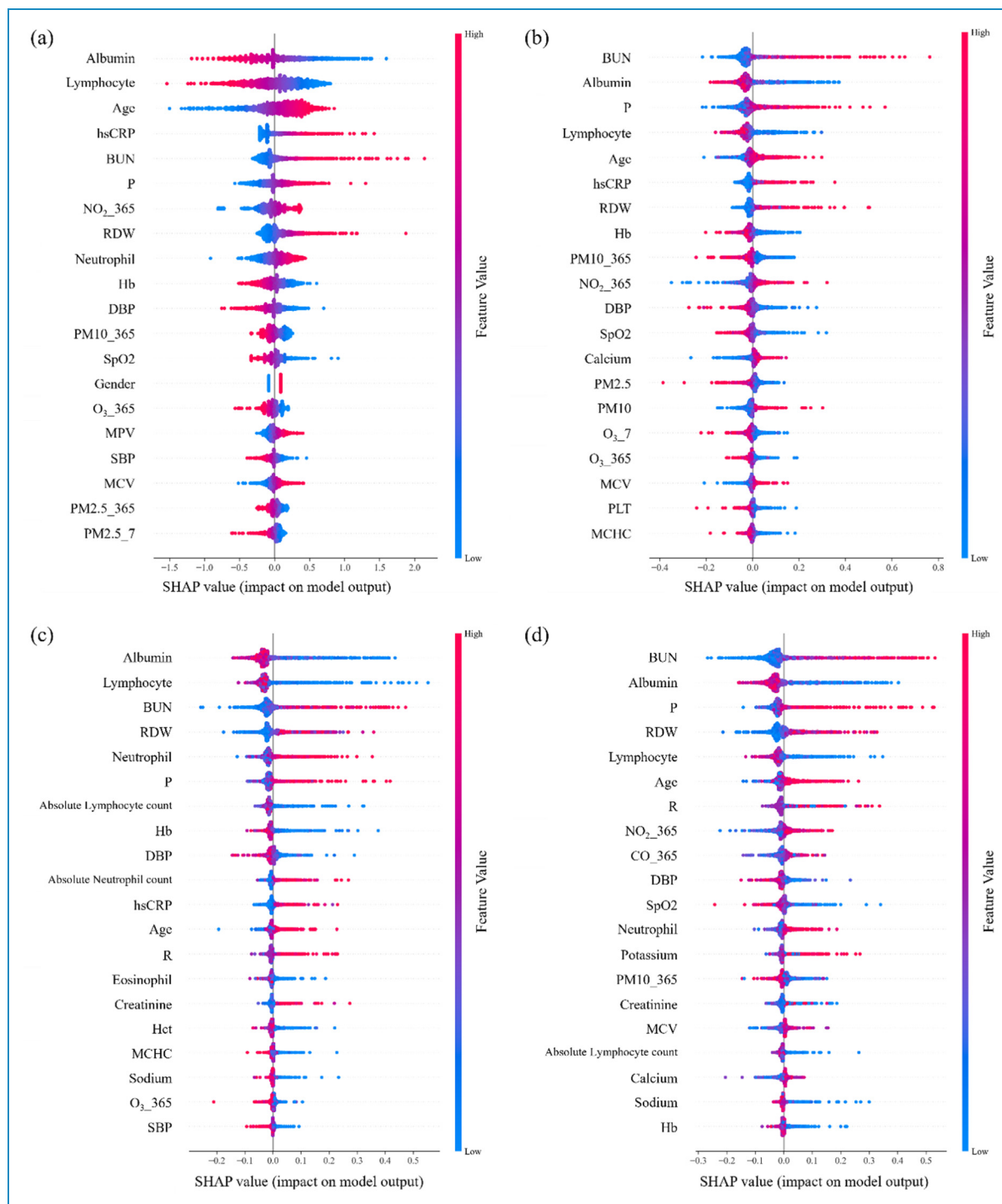


Figure 4. SHAP value based on feature importance for each prediction model, (a) LR, (b) SVM, (c) RF, (d) XGB, R: respiratory rate per min; P: pulse rate per min. LR: logistic regression; SVM: support vector machine; RF: random forest; XGB: extreme gradient boost; SHAP: Shapley additive explanations.

in various clinical settings. It is based on patient data recorded in the EMR, including demographic, clinical, and environmental characteristics as well as air pollution exposure levels obtained from publicly available online climate data. Additionally, clinical characteristics were

based on primary test items conducted in all medical institutions. The methodological characteristics of this study ensure the practical utility of the predictive model. Finally, the proposed model uses mathematical methods to present the relationships between the predictive

variables. The influence and directionality of each variable were quantitatively indicated using the permutation importance and SHAP values. This is significant because it interprets machine learning models that learn by considering complex interactions among independent variables in an explainable manner.

Nevertheless, this study has some limitations. First, the dataset used was insufficiently large. The dataset is highly unbalanced, with only an expired group size of 628 in the test data, which is relatively small compared with previous studies.⁸ Second, as this study was designed as a retrospective chart review study, it could not assess the severity of patients' conditions at the time of admission. Third, the model training used patient data from a specific region of South Korea. This implies that the predictive model reflects regional characteristics. Last, an external institution dataset was not constructed separately for model validation. The dataset from a single institution was split into training and test data, possibly leading to overestimating the test data prediction results. Future large-scale prospective studies should be designed to collect sufficient data, including patient severity, from various regional medical institutions to address these limitations.

Conclusion

The method proposed in this study predicted in-hospital mortality in patients with chronic respiratory diseases using demographic, clinical, and environmental characteristics. If this predictive model is introduced into clinical practice, it could enable early assessment of patient prognosis based on characteristics collected at admission, classifying patients into high- and low-risk groups. This is expected to assist healthcare professionals in clinical decision-making, enabling timely treatment and reducing the socioeconomic costs associated with chronic respiratory diseases.

Acknowledgements: This work was supported by the Gachon University research fund of 2023(GCU-202308020001). This work was supported by the GRRC program of Gyeonggi Province. [GRRC-Gachon2023(B01), Development of AI-based medical imaging technology]. This work was supported by the Technology Innovation Program(or Industrial Strategic Technology Development Program(K_G012001185601, Building Data Sets for Artificial Intelligence Learning) funded By the Ministry of Trade Industry & Energy(MOTIE, Korea).

Author contributions: Seung Yeob Ryu and Seon Min Lee contributed equally to this work. Seung Yeob Ryu and Seon Min Lee are the co-first (lead) authors.

Declaration of conflicting interests: The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding: The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Ministry of Trade Industry & Energy(MOTIE, Korea), GRRC program of Gyeonggi Province, (grant number K_G012001185601, GRRC-Gachon2023(B01)).

ORCID iDs: Seon Min  <https://orcid.org/0009-0004-8694-2169>
Young Jae Kim  <https://orcid.org/0000-0003-0443-0051>
Kwang Gi Kim  <https://orcid.org/0000-0001-9714-6038>

Supplemental material: Supplemental material for this article is available online.

Patient consent statement: This study was approved by the IRB as a retrospective medical records dataset collection study, and patient consent was not required.

References

1. Lopez K, Li H, Lipkin-Moore Z, et al. Deep learning prediction of hospital readmissions for asthma and COPD. *Respir Res* 2023; 24: 311. 20231213.
2. Iheanacho I, Zhang S, King D, et al. Economic burden of chronic obstructive pulmonary disease (COPD): a systematic literature review. *Int J Chron Obstruct Pulmon Dis* 2020; 15: 439–460. 20200226.
3. Ji Z, Li X, Lei S, et al. A pooled analysis of the risk prediction models for mortality in acute exacerbation of chronic obstructive pulmonary disease. *Clin Respir J* 2023; 17: 707–718. 20230321.
4. Molino NA, Turcatel G and Riskin D. Machine learning approaches to predict asthma exacerbations: a narrative review. *Adv Ther* 2024; 41: 534–552. 20231219.
5. Rezaee ME, Ward CE, Nunez B, et al. Examining 30-day COPD readmissions through the emergency department. *Int J Chron Obstruct Pulmon Dis* 2018; 13: 109–120. 20171227.
6. Singh D, Wedzicha JA, Siddiqui S, et al. Blood eosinophils as a biomarker of future COPD exacerbation risk: pooled data from 11 clinical trials. *Respir Res* 2020; 21: 240. 20200917.
7. Yoon EC, Koo SM, Park HY, et al. Predictive role of white blood cell differential count for the development of acute exacerbation in Korean chronic obstructive pulmonary disease. *Int J Chron Obstruct Pulmon Dis* 2024; 19: 17–31. 20240104.
8. Goto T, Camargo CA Jr., Faridi MK, et al. Machine learning approaches for predicting disposition of asthma and COPD exacerbations in the ED. *Am J Emerg Med* 2018; 36: 1650–1654. 20180628.
9. Shiroshita A, Kimura Y, Shiba H, et al. Predicting in-hospital death in pneumonic COPD exacerbation via BAP-65, CURB-65 and machine learning. *ERJ Open Res* 2022; 8: 20220124.
10. Zein JG, Wu CP, Attaway AH, et al. Novel machine learning can predict acute asthma exacerbation. *Chest* 2021; 159: 1747–1757. 20210110.
11. Xiong S, Chen W, Jia X, et al. Machine learning for prediction of asthma exacerbations among asthmatic patients: a systematic review and meta-analysis. *BMC Pulm Med* 2023; 23: 278. 20230728.

12. Poole JA, Barnes CS, Demain JG, et al. Impact of weather and climate change with indoor and outdoor air quality in asthma: a work group report of the AAAAI environmental exposure and respiratory health committee. *J Allergy Clin Immunol* 2019; 143: 1702–1710. 20190228.
13. Tiotiu AI, Novakova P, Nedeva D, et al. Impact of air pollution on asthma outcomes. *Int J Environ Res Public Health* 2020; 17: 20200827.
14. Ryu B, Yoo S, Kim S, et al. Thirty-day hospital readmission prediction model based on common data model with weather and air quality data. *Sci Rep* 2021; 11: 23313. 20211202.
15. Byrd JB, Newby DE, Anderson JA, et al. Blood pressure, heart rate, and mortality in chronic obstructive pulmonary disease: the SUMMIT trial. *Eur Heart J* 2018; 39: 3128–3134.
16. Elvekjaer M, Aasvang EK, Olsen RM, et al. Physiological abnormalities in patients admitted with acute exacerbation of COPD: an observational study with continuous monitoring. *J Clin Monit Comput* 2020; 34: 1051–1060. 20191111.
17. Hoepers AT, Menezes MM and Frode TS. Systematic review of anaemia and inflammatory markers in chronic obstructive pulmonary disease. *Clin Exp Pharmacol Physiol* 2015; 42: 231–239.
18. Vedel-Krogh S, Fallgaard Nielsen S, Lange P, et al. Association of blood eosinophil and blood neutrophil counts with asthma exacerbations in the Copenhagen general population study. *Clin Chem* 2017; 63: 823–832. 20170216.
19. Semenzato U, Biondini D, Bazzan E, et al. Low-blood lymphocyte number and lymphocyte decline as key factors in COPD outcomes: a longitudinal cohort study. *Respiration* 2021; 100: 618–630. 20210426.
20. Wang M, Zhang J, Ji Q, et al. Evaluation of platelet distribution width in chronic obstructive pulmonary disease patients with pulmonary embolism. *Biomark Med* 2016; 10: 587–596. 20151116.
21. Alparslan Bekir S, Tuncay E, Gungor S, et al. Can red blood cell distribution width (RDW) level predict the severity of acute exacerbation of chronic obstructive pulmonary disease (AECOPD)? *Int J Clin Pract* 2021; 75: e14730. 20210819.
22. Seyhan EC, Ozgul MA, Tutar N, et al. Red blood cell distribution and survival in patients with chronic obstructive pulmonary disease. *COPD* 2013; 10: 416–424. 20130328.
23. Tertemiz KC, Ozgen Alpaydin A, Sevinc C, et al. Could “red cell distribution width” predict COPD severity? *Rev Port Pneumol (2006)* 2016; 22: 196–201. 20160122.
24. Chalela R, Gonzalez-Garcia JG, Chillaron JJ, et al. Impact of hyponatremia on mortality and morbidity in patients with COPD exacerbations. *Respir Med* 2016; 117: 237–242. 20160629.
25. Wan X, Chen L, Zhu Z, et al. Association of Serum calcium with the risk of chronic obstructive pulmonary disease: a prospective study from UK biobank. *Nutrients* 2023; 15: 20230803.
26. Chen L, Chen L, Zheng H, et al. The association of blood urea nitrogen levels upon emergency admission with mortality in acute exacerbation of chronic obstructive pulmonary disease. *Chron Respir Dis* 2021; 18: 14799731211060051.
27. Zhang J, Qin Y, Zhou C, et al. Elevated BUN upon admission as a predictor of in-hospital mortality among patients with acute exacerbation of COPD: a secondary analysis of multicenter cohort study. *Int J Chron Obstruct Pulmon Dis* 2023; 18: 1445–1455. 20230713.
28. Zinellu E, Fois AG, Sotgiu E, et al. Serum albumin concentrations in stable chronic obstructive pulmonary disease: a systematic review and meta-analysis. *J Clin Med* 2021; 10: 20210113.
29. Ling M, Huiyin L, Shanglin C, et al. Relationship between human serum albumin and in-hospital mortality in critical care patients with chronic obstructive pulmonary disease. *Front Med (Lausanne)* 2023; 10: 1109910. 20230427.
30. Blakey JD, Price DB, Pizzichini E, et al. Identifying risk of future asthma attacks using UK medical record data: a respiratory effectiveness group initiative. *J Allergy Clin Immunol Pract* 2017; 5: 1015–1024 e1018. 20161222.
31. Steinwart I and Christmann A. *Support vector machines*. 1st ed. New York: Springer, 2008, p.xvi, 601 p.
32. Vapnik VN. *The nature of statistical learning theory*. 2nd ed. New York: Springer, 2000, p.xix, 314 p.
33. Breiman L. Random forests. *Mach Learn* 2001; 45: 5–32.
34. Carreira-Perpiñán MÁ and Zharmagambetov A. Ensembles of bagged TAO trees consistently improve over random forests, AdaBoost and gradient boosting. Proceedings of the 2020 ACM-IMS on Foundations of Data Science Conference. Virtual Event, USA: Association for Computing Machinery, 2020, p. 35–46.
35. Wade C and Glynn K. *Hands-On Gradient Boosting with XGBoost and Scikit-learn: Perform Accessible Machine Learning and Extreme Gradient Boosting with Python*. 1st ed. Birmingham: Packt Publishing, Limited, 2020.
36. Fisher A, Rudin C and Dominici F. All models are wrong, but many are useful: learning a variable’s importance by studying an entire class of prediction models simultaneously. *J Mach Learn Res* 2019; 20: 1–18.
37. Aas K, Jullum M and Løland A. Explaining individual predictions when features are dependent: more accurate approximations to Shapley values. *Artif Intell* 2021; 298: 103502.
38. Lundberg SM and Lee SI. A unified approach to interpreting model predictions. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, California, USA: Curran Associates Inc., 2017, p. 4768–4777.
39. Shengping Y and Gilbert B. The receiver operating characteristic (ROC) curve. *The Southwest Respiratory Crit Care Chronicles* 2017; 5: 34–36.
40. Halper-Stromberg E, Yun JH, Parker MM, et al. Systemic markers of adaptive and innate immunity are associated with chronic obstructive pulmonary disease severity and spirometric disease progression. *Am J Respir Cell Mol Biol* 2018; 58: 500–509.
41. Zhou C, Yi Q, Luo Y, et al. Low diastolic blood pressure and adverse outcomes in inpatients with acute exacerbation of chronic obstructive pulmonary disease: a multicenter cohort study. *Chin Med J (Engl)* 2023; 136: 941–950. 20030407.
42. Agnello L, Giglio RV, Bivona G, et al. The value of a complete blood count (CBC) for sepsis diagnosis and prognosis. *Diagnostics (Basel)* 2021; 11: 20211012.
43. Danesh Yazdi M, Wang Y, Di Q, et al. Long-term exposure to PM(2.5) and ozone and hospital admissions of medicare participants in the southeast USA. *Environ Int* 2019; 130: 104879. 20190622.