Journal of Asthma and Allergy

Dovepress
open access to scientific and medical research

Open Access Full Text Article

SHORT REPORT

Predicting Asthma Exacerbation Risk in the Adult South Korean Population Using Integrated Health Data and Machine Learning Models

Joon Young Choi[1], Chin Kook Rhee[2]

[1]Department of Internal Medicine, Incheon St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Seoul, Republic of Korea; [2]Department of Internal Medicine, Seoul St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Seoul, Republic of Korea

Correspondence: Chin Kook Rhee, Division of Pulmonary and Critical Care Medicine, Department of Internal Medicine, Seoul St. Mary's Hospital, College of Medicine, The Catholic University of Korea, 222 Banpo-daero, Seocho-gu, Seoul, 06591, Republic of Korea, Tel +82 2 2258 6067, Fax +82 2 599 3589, Email chinkook77@gmail.com

Abstract: Asthma is a chronic inflammatory airway disease with significant burden; exacerbations can severely affect quality of life and healthcare costs. Advances in big data analysis and artificial intelligence have made it easier to predict future exacerbations more accurately. This study used an integrated dataset of Korean National Health Insurance, meteorological, air pollution, and viral data from national public databases to develop a model to predict asthma exacerbations on a daily basis in South Korea. We merged these sources and applied random forest, AdaBoost, XGBoost, and LightGBM machine learning models to compare their performances at predicting future exacerbations. Of the models, XGBoost (AUROC of 0.68 and accuracy of 0.96) and LightGBM (AUROC of 0.67 and accuracy of 0.96) were the most promising. Common important variables were the number of visits and exacerbations per year, and medical resource utilization, including the prescription of asthma medications. Comorbid diabetes, hypertension, gastroesophageal reflux, arthritis, metabolic syndrome, osteoporosis, and ischemic heart disease were also associated with elevated exacerbation risk. The models examined in this study highlight the importance of previous exacerbations, use of medical resources, and comorbidities in the prediction of future exacerbations in patients with asthma.

Keywords: asthma, machine learning, big data analysis, South Korea, XGBoost, LightGBM

Introduction

Asthma is a chronic respiratory condition characterized by airway inflammation, hyper-responsiveness, and remodeling that poses a significant public health burden globally.[1] Exacerbation, acute worsening episodes of asthma symptoms, not only impairs quality of life but also substantially contributes to healthcare costs and resource utilization.[2] Predicting future exacerbations may lead to preemptive interventions, improve patient outcomes, and reduce healthcare burden.

Previous predictions of exacerbation have been based mostly on clinical parameters and patient-reported outcomes, using parametric methods. While valuable, these models did not capture the full spectrum of factors influencing exacerbation risk. Specifically, they mostly did not integrate comprehensive clinical, environmental, and viral data, which are crucial for a holistic understanding of factors associated with exacerbation. Furthermore, traditional parametric methods used in earlier studies may not effectively handle the complex interactions between multiple variables, leading to less accurate predictions. Previous predictive model for exacerbation was on an annual basis. There have been few models that predicts risk of exacerbation on a daily basis. However, advances in big data analysis and artificial intelligence models have the potential to make more accurate and timely predictions.[3–7] In this study, we integrated national health insurance data with viral, meteorological, and air pollution data to develop a daily predictive model for asthma exacerbations.

Received: 2 April 2024
Accepted: 8 August 2024
Published: 13 August 2024

Methods

Study Design and Data Sources

We merged data from four sources: the National Health Insurance Service-National Sample Cohort Database (NHIS-NSC), viral data from the Korean Centers for Disease Control (KDCA), meteorological data from the Korea Meteorological Administration (KMA), and air pollution data from Air Korea.

NHIS-NSC

NHIS-NSC is a nationwide population-based sample cohort that contains data from 2002 to 2015 on 1,025,340 individuals (2.2% of the total population in 2002) who were selected based on the entire population maintaining national health insurance or medical aid in the year 2006. The database includes demographics, socioeconomic status, health insurance claim codes, diagnostic codes, and medical examination data. For this study, the medical data of patients with asthma were extracted from the database between 2008 and 2013 based on the following criteria: age ≥ 15 years; 10th International Statistical Classification of Disease and Related Health Problems (ICD-10) code for asthma (J45.x-J46.x) in principle or within the fourth position of secondary diagnosis; and more than one prescription for asthma medication in the previous year, including an inhaled corticosteroid (ICS), ICS plus long-acting β2-agonists (LABA), leukotriene receptor antagonists (LTRA), short-acting β2-agonists (SABA), short-acting muscarinic antagonists (SAMA), SAMA+SABA, systemic corticosteroids, methylxanthine, or systemic bronchodilators.

Viral Data from the KDCA

Since 2000, the KDCA has implemented the Korea Influenza and Respiratory Viruses Surveillance System for ongoing monitoring of respiratory viruses across Korea. This surveillance framework has specimen-based and clinical surveillance system components. The specimens were collected from patients who visited outpatient clinics with acute respiratory illnesses at 52 designated sentinel sites; clinical information was collected from patients who were hospitalized. This comprehensive surveillance system provides weekly updates on the detection and positivity rates of eight respiratory viruses (influenza virus, parainfluenza virus, adenovirus, respiratory syncytial virus, coronavirus, rhinovirus, bocavirus, and enterovirus) identified by multiplex reverse transcription-polymerase chain reaction (RT-PCR) analysis. These data are publicly available on the KDCA website (http://www.kdca.go.kr/npt/).

Meteorological Data from the KMA

We merged meteorological data provided by the KMA, which is publicly available at https://data.kma.go.kr/data/grnd/selectAsosRltmList.do?pgmNo=36; the data include average, minimum, and maximum temperatures, average, and maximum wind speeds, total wind run, average dew point temperature, daily precipitation, minimum, and average relative humidity, total sunshine hours, total solar radiation, and average local atmospheric pressure. Given the presence of multiple observation stations within each province, we calculated daily summary statistics for each province by averaging the daily observations for all stations within that province to obtain provincial daily average observations.

Air Pollution Data from Air Korea

Air Korea is a comprehensive real-time air quality information system launched in 2005, developed by Korea Environment Corporation. It uses data from 642 (or more) monitoring stations nationwide, including urban, suburban, roadside, and port air monitoring networks, providing data on air pollutants such as sulfur dioxide (SO_2), carbon monoxide (CO), nitrogen dioxide (NO_2), ozone (O_3), and particulate matter (PM_{10} and $PM_{2.5}$). We extracted daily average concentrations of SO_2, CO, O_3, NO_2, and PM_{10} for 16 provinces in South Korea. For each region, we calculated the daily mean value of observations from multiple monitoring stations. This study did not include data on $PM_{2.5}$ due to a lack of availability.

Definition of Exacerbation

Operationally, a moderate exacerbation was defined as an outpatient claim that includes asthma (J45.x-J46.x) in principle or within the fourth position of secondary diagnosis of ICD-10 codes, and a concurrent claim for systemic corticosteroids or antibiotics. Severe exacerbation was defined as a claim for an admission or emergency room visit, with asthma (J45.x-J46.x) or associated codes (R06.0, J80) in principle or within the fourth position of secondary diagnoses of ICD-10 codes, and a concurrent claim for systemic steroids or antibiotics.

Statistical Analysis

Data Splitting

To train and evaluate the performance of the exacerbation prediction models, the entire dataset was divided into training and test data at a 70:30 ratio. The training data were used to train the prediction models and the test data were used to evaluate the performance of the trained model. Of the 168,730 subjects with asthma, 11,811 were allocated to the training data and 50,619 to the test data, with respective exacerbation rates of 2.41% and 2.42% in each dataset.

Imbalanced Data Handling

Due to the significant imbalance in exacerbation rates, sampling methods were used to balance the data. To improve the accuracy of the prediction model, initially only under-sampling was applied to the training data to adjust the ratio of exacerbation occurrence to non-occurrence to 88:12. Subsequently, a hybrid sampling approach combining under-sampling and adaptive synthetic sampling was used to adjust the ratio to 67:33 in the training data.

Variable Selection

After merging four databases, the total dataset included 151 variables (Table 1). Redundant or excessive variables may lead to a complex model prone to overfitting, resulting in impaired generalization and accuracy. To address this issue, we used tree-based feature selection to identify 28 critical variables based on their importance.

Table 1 Variables for the Prediction Model

Variables	
Basic characteristics	**Sex, age**, visit day, visit week, visit year, **number of claims per year, number of visit per year**
Location	Province
Exacerbation	Mild/severe/any exacerbation, number of mild/severe/**any exacerbation per year**
Asthma code	**Number of asthma code**, or asthma-like code per year, Asthma code or asthma-like code
Antibiotics/hospitalization/ ER visit	**Number of antibiotics prescription per year, number of ER visit or hospitalization per year**, antibiotics prescription, ER visit or hospitalization
Comorbidities code per year	**Ischemic heart disease**, lung cancer, **osteoporosis, depression, arthritis, DM, GERD**, pneumonia, heart failure, **hypertension**, anemia, **metabolic syndrome**
Asthma medication prescription per year	**IV corticosteroid, oral corticosteroid, systemic bronchodilator, methylxanthine, ICS, ICS/LABA**, LABA, LAMA, LABA/LAMA, SAMA, **SABA**, SABA/SAMA, **LTRA**
General health screening data	Date of screening, **BMI, systolic/diastolic BP, total cholesterol**, smoking status, smoking duration, smoking consumption per day
Meteorological data	Minimal temperature, daily precipitation, maximum/average wind speed, total wind run, average dew point temperature, minimum/average relative humidity, average local atmospheric pressure, total sunshine hours, total solar radiation, daily temperature range, minimum temperature previous day, lowest temperature accumulated over 3/5/7 days, lowest temperature difference for 3/5/7 days, accumulated precipitation for 3/5/7 days,
Viral data	Total virus detection, total virus detection rate, adenovirus/bocavirus/coronavirus/enterovirus/human metapneumovirus/rhinovirus/influenza virus/ parainfluenza virus/RSV detection rate for the week, adenovirus/ bocavirus/coronavirus/enterovirus/human metapneumovirus/rhinovirus/influenza virus/ parainfluenza virus/RSV detection rate in the previous week, sum of 2-4-week adenovirus/bocavirus/coronavirus/enterovirus/human metapneumovirus/rhinovirus/influenza virus/ parainfluenza virus/RSV detection rate,
Air pollution data	CO/NO_2/O_3/PM10/SO_2 concentration, CO/NO_2/O_3/PM10/SO_2 concentration in the previous day, sum of 3-5-7-day CO/NO_2/O_3/PM10/SO_2 concentration

Notes: Twenty-eight critical variables identified by tree-based feature selection is identified as bold character.

Prediction Model Development

To address the severe class imbalance issue in the training data, models were fit using data resampled by under-sampling and hybrid sampling, as well as data solely adjusted by under-sampling. Models were fit using both the 28 explanatory variables selected through variable selection and the full set of 151 explanatory variables. The models included parametric methods such as logistic regression and penalized regression, as well as random forest, AdaBoost, XGBoost, and LightGBM machine learning methods. The metrics used to evaluate the models were the area under the receiver operating characteristic (AUROC) curve, accuracy, precision, recall, and F1 score.

Results

We identified 168,730 patients with asthma using the operational definition. Of these, 10,384 (6.15%) had experienced at least one exacerbation in the 6-year study period. The mean exacerbation frequency in the study period was 3.71; the mean frequency of severe exacerbations was 0.15.

Daily Prediction Model of Exacerbation in Asthma Patients Using Test Data

Table 2 compares model performance using test data according to five evaluation metrics. Overall, the model with superior predictive performance addressed the class imbalance issue using under-sampling and applied the XGBoost model to the 156 explanatory variables. It had an AUROC of 0.68 and accuracy of 0.96. The LightGBM model with under-sampling using 156 variables also had excellent performance.

Variable Importance

To identify key variables influencing the prediction of acute asthma exacerbations, we used ensemble models based on classification trees. Of the models, we focused on those with good performance metrics with the training data (XGBoost and LightGBM), examining the top 25 variables in terms of importance.

Common important explanatory variables identified by both XGBoost and LightGBM for predicting acute exacerbations included age, systolic, and diastolic blood pressure, number of disease codes for asthma, metabolic syndrome, arthritis, hypertension, and gastroesophageal reflux disease annually, and the annual numbers of visits, exacerbations, claims, ER visits or hospitalizations, and asthma medication prescriptions, including systemic bronchodilators, IV, and oral corticosteroid, LTRA, SABA, and methylxanthine (Figures 1 and 2). In both models, the number of visits and exacerbations per year were the most important factors associated with future exacerbations (XGBoost: F scores of 235 and 201, respectively; LightGBM: F scores of 290 and 682, respectively).

Table 2 Model Evaluation Results Using Test Data

Method	Sampling Method	Feature Selection	AUC	Precision	Recall	F1-score	Accuracy
1) Logistic regression	Under	No	0.59	0.31	0.18	0.23	0.97
2) penalized logistic regression (L1)	Under	No	0.59	0.31	0.19	0.23	0.97
3) penalized logistic regression (L1)	Hybrid	Yes	0.66	0.22	0.36	0.27	0.95
4) Random Forest	Hybrid	No	0.67	0.19	0.38	0.25	0.95
5) Random Forest	Hybrid	Yes	0.71	0.19	0.48	0.28	0.94
6) Adaboost	Under	No	0.64	0.27	0.31	0.29	0.96
7) Adaboost	Hybrid	No	0.70	0.17	0.46	0.25	0.93
8) Adaboost	Hybrid	Yes	0.71	0.17	0.48	0.25	0.93
9) light GBM	Under	No	0.67	0.28	0.37	0.32	0.96
10) light GBM	Hybrid	Yes	0.69	0.25	0.41	0.31	0.96
11) XGBoost	Under	No	0.68	0.28	0.38	0.32	0.96
12) XGBoost	Hybrid	No	0.68	0.27	0.38	0.32	0.96
13) XGBoost	Hybrid	Yes	0.68	0.26	0.39	0.31	0.96

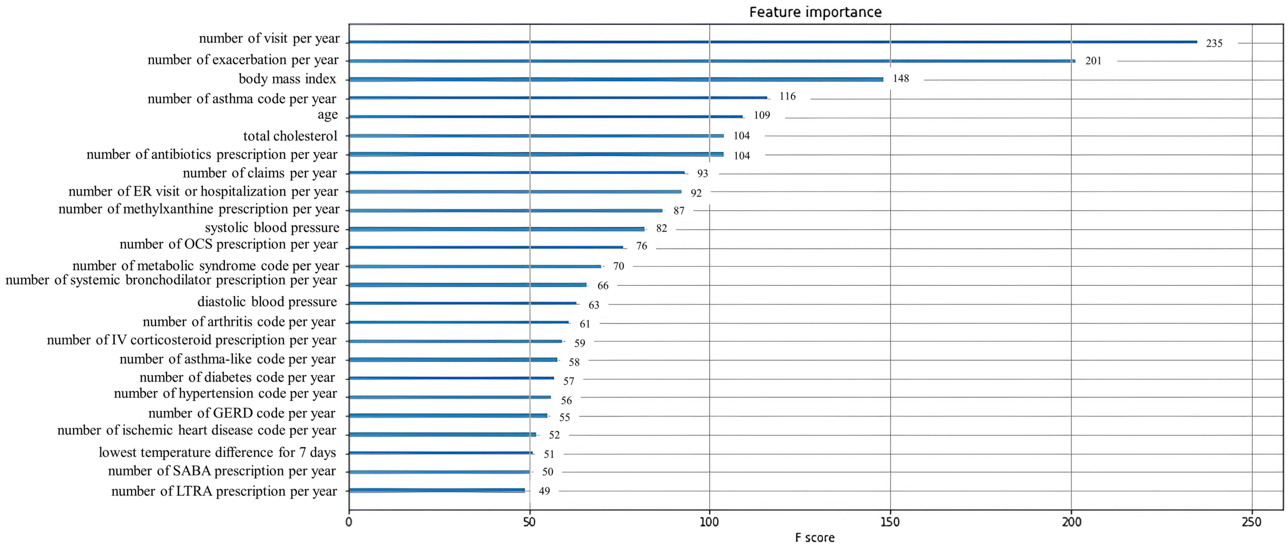

Figure 1 Explanatory variables identified by both XGBoost for predicting acute exacerbations.

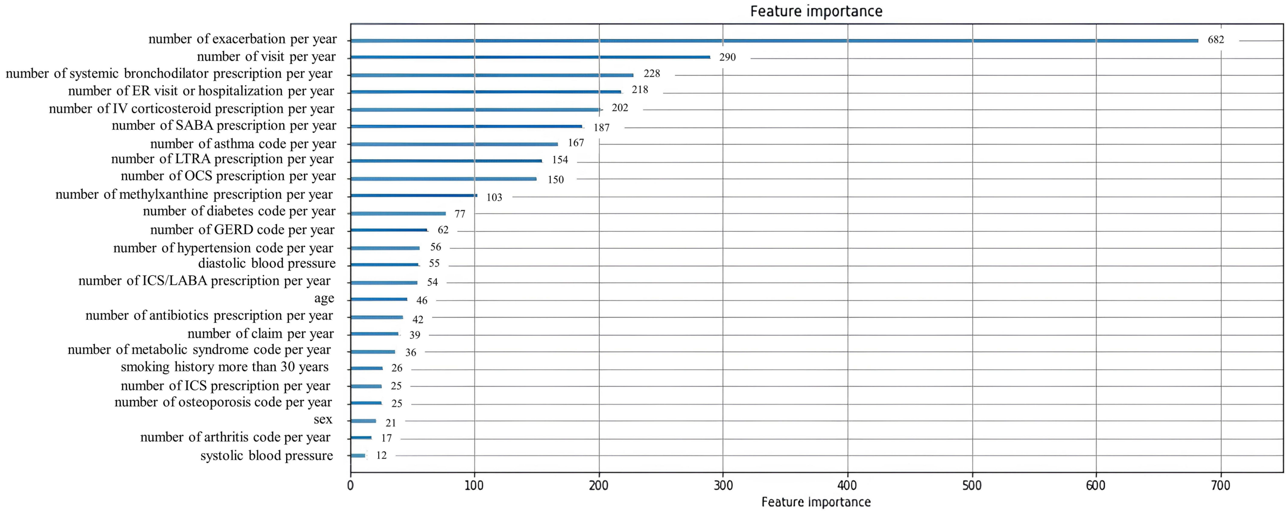

Figure 2 Explanatory variables identified by LightGBM for predicting acute exacerbations.

Discussion

We examined the performances of various models in predicting asthma exacerbation, using comprehensive integrated data from nationwide health records and environmental and viral data. XGBoost and LightGBM had the most promising performances. Important variables common to the two models were the numbers of visits and exacerbations per year. Medical resource utilization, including prescription of asthma medication, ER visits, and hospitalization were also important variables in both models. The comorbidities including diabetes, hypertension, gastroesophageal reflux, arthritis, metabolic syndrome, osteoporosis, and ischemic heart disease were also associated with elevated exacerbation risks.

Our results concur with previous studies that have shown strong predictive values of previous exacerbations and medication use for predicting future exacerbations.[3,8–10] Grossman et al[10] analyzed 1840 patients enrolled in 12 Asthma Clinical Research Network Trials and AsthmaNET trials to measure exacerbation risk in asthma patients. For the entire population, a history of exacerbation and percent predicted FEV_1 were associated with increased exacerbation rates (relative rates [RR] =1.9 [95% CI, 1.37–2.56] and 0.74 [95% CI, 0.63–0.87], respectively). However, in our predictive models, environmental (except temperature) and viral data had little impact on future exacerbation risk. Although previous studies have reported a negative impact of air pollution,[5,11–13] meteorological factors,[5,12,13] and viral exposure,[14] the impact of these factors may be trivial and masked by the previous exacerbation frequency and medical utilization. Further research is needed to explore whether these factors play a significant role under specific conditions or in conjunction with other variables. There are limitations in this study. First, lower airway tract infection is a potential risk factor for exacerbation. However, we did not include infection in our machine learning model. Second, adherence to inhaler is an important factor associated with exacerbation. However, due to the technical difficulty, we could not include adherence in our model.

In conclusion, our study highlights the potential of using integrated health data for predicting asthma exacerbations. Using various machine learning models, we identified the impact of previous asthma exacerbations and use of medical resources for predicting asthma exacerbations. Comorbid conditions such as diabetes, hypertension, gastroesophageal reflux disease, arthritis, metabolic syndrome, osteoporosis, and ischemic heart disease were also associated with an increased risk of exacerbations. In comparison, other environmental and viral factors have relatively little impact on the prediction models. Additionally, our results demonstrated the capability of advanced machine learning models to provide comprehensive risk assessments. Our study offers important results that can enhance clinical prediction and prevention of asthma exacerbations, leading to improved patient management and outcomes.

Abbreviations

NHIS-NSC, National Health Insurance Service-National Sample Cohort Database; KDCA, the Korean Centers for Disease Control; KMA, Korea Meteorological Administration; ICD-10, 10th International Statistical Classification of Disease and Related Health Problems; ICS, inhaled corticosteroid; LABA, long-acting β2-agonists; LTRA, leukotriene receptor antagonist; SABA, short-acting β2-agonists; SAMA, short-acting muscarinic antagonists; RT-PCR, reverse transcription-polymerase chain reaction; SO2, sulfur dioxide; CO, carbon monoxide; NO2, nitrogen dioxide; O3, ozone; PM, particulate matter; AUC, Area Under the ROC Curve; GERD, gastroesophageal reflux disease.

Data Sharing Statement

The datasets supporting the conclusions of this article are available from the corresponding author on reasonable request.

Ethics Approval and Consent to Participate

This study was approved by the Institutional Review Board of Seoul St. Mary's Hospital, Republic of Korea with registration number KC18ZNSI0850. Informed consent from patients was waived by the Board. Data accessed complied with relevant data protection and privacy regulations.

Author Contributions

All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal which the article has been submitted; and agree to be accountable for all aspects of the work.

Funding

This study was supported by grants from the Korean Environment Industry and Technology Institute through the Core Technology Development Project for Environmental Disease Prevention and Management, funded by the Korea Ministry of Environment (Grant number 2022003310008).

Disclosure

The authors report no conflicts of interest in this work.

References

1. Momtazmanesh S, Moghaddam SS, Ghamari S-H.; GBD 2019 Chronic Respiratory Diseases Collaborators. Global burden of chronic respiratory diseases and risk factors, 1990–2019: an update from the global burden of disease study 2019. *EClinicalMedicine*. 2023;59:101936. doi:10.1016/j.eclinm.2023.101936
2. Wisnivesky J, Federmann E, Eckert L, et al. Impact of exacerbations on lung function, resource utilization, and productivity: results from an observational, prospective study in adults with uncontrolled asthma. *J Asthma*. 2023;60:1072–1079. doi:10.1080/02770903.2022.2130800
3. Jiao T, Schnitzer ME, Forget A, Blais L. Identifying asthma patients at high risk of exacerbation in a routine visit: a machine learning model. *Respir Med*. 2022;198:106866. doi:10.1016/j.rmed.2022.106866
4. Joo H, Lee D, Lee SH, Kim YK, Rhee CK. Increasing the accuracy of the asthma diagnosis using an operational definition for asthma and a machine learning method. *BMC Pulm Med*. 2023;23:196. doi:10.1186/s12890-023-02479-4
5. Hwang H, Jang JH, Lee E, Park HS, Lee JY. Prediction of the number of asthma patients using environmental factors based on deep learning algorithms. *Respir Res*. 2023;24:302. doi:10.1186/s12931-023-02616-x
6. Jo YS, Han S, Lee D, et al. Development of a daily predictive model for the exacerbation of chronic obstructive pulmonary disease. *Sci Rep*. 2023;13:18669. doi:10.1038/s41598-023-45835-4
7. Lee J, Jung HM, Kim SK, et al. Factors associated with chronic obstructive pulmonary disease exacerbation, based on big data analysis. *Sci Rep*. 2019;9:6679. doi:10.1038/s41598-019-43167-w
8. He S, Lin W, Zhong J, Zheng X, Jin Y, Cao C. Independent risk factors of asthma exacerbations: 3-year follow-up in a single-center prospective cohort study. *Ann Transl Med*. 2022;10(24):1353. doi:10.21037/atm-22-5918
9. Kraft M, Brusselle G, FitzGerald JM, et al. Patient characteristics, biomarkers and exacerbation risk in severe, uncontrolled asthma. *Euro Respir J*. 2021;2021:58.
10. Grossman NL, Ortega VE, King TS, et al. Exacerbation-prone asthma in the context of race and ancestry in Asthma clinical research network trials. *J Allergy Clin Immunol*. 2019;144:1524–1533. doi:10.1016/j.jaci.2019.08.033
11. Tiotiu AI, Novakova P, Nedeva D, et al. Impact of air pollution on asthma outcomes. *Int J Environ Res Public Health*. 2020;18:17. doi:10.3390/ijerph18010017
12. Yu HR, Lin CR, Tsai JH, et al. A multifactorial evaluation of the effects of air pollution and meteorological factors on asthma exacerbation. *Int J Environ Res Public Health*. 2020;18:17.
13. Jo EJ, Choi MH, Kim CH, et al. Patterns of medical care utilization according to environmental factors in asthma and chronic obstructive pulmonary disease patients. *Korean J Intern Med*. 2021;36:1146–1156. doi:10.3904/kjim.2020.168
14. Mikhail I, Grayson MH. Asthma and viral infections: an intricate relationship. *Ann Allergy Asthma Immunol*. 2019;123:352–358. doi:10.1016/j.anai.2019.06.020