Original Paper

Obtaining the Most Accurate, Explainable Model for Predicting Chronic Obstructive Pulmonary Disease: Triangulation of Multiple Linear Regression and Machine Learning Methods

Arnold Kamis*, PhD; Nidhi Gadia*, MS; Zilin Luo*, MS; Shu Xin Ng*, MS; Mansi Thumbar*, MS

Brandeis International Business School, Brandeis University, Waltham, MA, United States

*all authors contributed equally

Corresponding Author:
Arnold Kamis, PhD
Brandeis International Business School
Brandeis University
Sachar International Center
415 South St
Waltham, MA, 02453
United States
Phone: 1 781 736 8544
Fax: 1 781 736 2269
Email: akamis@brandeis.edu

Abstract

Background: Lung disease is a severe problem in the United States. Despite the decreasing rates of cigarette smoking, chronic obstructive pulmonary disease (COPD) continues to be a health burden in the United States. In this paper, we focus on COPD in the United States from 2016 to 2019.

Objective: We gathered a diverse set of non–personally identifiable information from public data sources to better understand and predict COPD rates at the core-based statistical area (CBSA) level in the United States. Our objective was to compare linear models with machine learning models to obtain the most accurate and interpretable model of COPD.

Methods: We integrated non–personally identifiable information from multiple Centers for Disease Control and Prevention sources and used them to analyze COPD with different types of methods. We included cigarette smoking, a well-known contributing factor, and race/ethnicity because health disparities among different races and ethnicities in the United States are also well known. The models also included the air quality index, education, employment, and economic variables. We fitted models with both multiple linear regression and machine learning methods.

Results: The most accurate multiple linear regression model has variance explained of 81.1%, mean absolute error of 0.591, and symmetric mean absolute percentage error of 9.666. The most accurate machine learning model has variance explained of 85.7%, mean absolute error of 0.456, and symmetric mean absolute percentage error of 6.956. Overall, cigarette smoking and household income are the strongest predictor variables. Moderately strong predictors include education level and unemployment level, as well as American Indian or Alaska Native, Black, and Hispanic population percentages, all measured at the CBSA level.

Conclusions: This research highlights the importance of using diverse data sources as well as multiple methods to understand and predict COPD. The most accurate model was a gradient boosted tree, which captured nonlinearities in a model whose accuracy is superior to the best multiple linear regression. Our interpretable models suggest ways that individual predictor variables can be used in tailored interventions aimed at decreasing COPD rates in specific demographic and ethnographic communities. Gaps in understanding the health impacts of poor air quality, particularly in relation to climate change, suggest a need for further research to design interventions and improve public health.

(JMIR AI 2024;3:e58455) doi: 10.2196/58455

KEYWORDS

chronic obstructive pulmonary disease; COPD; cigarette smoking; ethnic and racial differences; machine learning; multiple linear regression; household income; practical model

Introduction

Background

Lung disease is a severe problem in the United States. According to the Centers for Disease Control and Prevention (CDC), asthma is responsible for at least 3000 deaths per year and chronic obstructive pulmonary disease (COPD) is responsible for at least 150,000 deaths per year. COPD is a progressive lung disease, encompassing chronic bronchitis and emphysema, which is characterized by airflow limitation and breathing difficulties. Asthma and COPD can co-occur (asthma-COPD overlap), with increased risk of mortality [1] and diminished disease-related quality of life [2]. This is from a variety of factors, some under individual control, such as cigarette smoking, and others not under individual control, such as ambient air pollution.

Cigarette smoking has been trending downward in recent years, thanks in part to public health advertisement campaigns. Nevertheless, air quality can be dangerously poor at times, which exacerbates lung health problems [3], and the impacts can be particularly acute in populations considered vulnerable. Technologically, there are tools that help individuals avoid poor air quality. For example, there are mobile phone apps that track air quality. They notify their owners on days when air quality is dangerously poor, advising them to stay indoors or avoid strenuous outdoor exercise. The effectiveness of such apps is ambiguous thus far [4,5].

The rest of the paper is organized as follows. We first review prior work regarding the possible factors contributing to COPD in adults. We then describe our methods, including data sources for the variables of interest and descriptive statistics. Following this, we will describe and interpret the results of our multiple linear regression (MLR) and machine learning (ML) models. We conclude by describing the overall research contributions as well as limitations and future directions.

Prior Work

There is substantial literature on factors contributing to COPD, including a wide variety of environmental, economic, and demographic variables; the etiology of COPD is multifactorial, with smoking being the most well-known contributing factor. Furthermore, the combination of environmental pollutants and cigarette smoke has shown synergistic effects, accelerating the decline in lung function and worsening COPD [6,7]. In addition, occupational exposures, for example, to coal dust, arsenic, or diesel fumes, or to home exposures, such as gas stoves, wood stoves, kerosene heaters, and fireplaces, contribute to overall COPD outcomes. When combined with persistent ambient air pollution, the risk and severity of COPD will likely increase [8].

Pollutants and copollutants are associated with decreased lung function and can lead to COPD. The loss can range from mild, such as allergies, to severe, that is, mortality. Air quality varies widely throughout the United States because of pollutants and copollutants, and climate change may be worsening it, particularly for populations considered vulnerable [9]. Health disparities due to poor quality air and other stressors are well known [10-12]. Ambient air pollution in poorer neighborhoods tend to be exacerbated by additional copollutants, heat stress, and aeroallergens. Air quality index (AQI) includes the totality of pollutants and copollutants.

ML methods have been applied increasingly to public health and medical problems. For example, ML has been used to support the public health response to COVID-19 through surveillance, case identification, contact tracing, and evaluating interventions [13]. ML methods have been used as a supportive tool to recognize cardiac arrest in emergency calls [14]. In that study, Zicari et al [14] developed a general protocol with a collaborative team to ensure that the ML tool was domain- and context-sensitive as well as abiding by ethical guidelines, thus obtaining trustworthiness. ML has been also used to improve early and accurate stroke recognition during emergency medical calls [15].

ML methods have been used to study COPD, in particular. For example, ML methods have been used to develop a prediction system using lifestyle data, environmental factors, and patient symptoms for the early detection of acute exacerbations of COPD within a 7-day window [16]. Another study on acute exacerbations of COPD compared several ML methods and found that a decision tree classifier was best for assessing patient severity and guiding treatment strategy [17]. In another study, to improve mortality prediction from COPD, a random forest was used to identify the most important imaging features [18]. Gradient boosted trees (GBTs) have been used to predict lung function values from computed tomography images obtained from patients with COPD and those without COPD [19]. Deep learning has been effective in analyzing images diagnostic of COPD [20]. Finally, research using a generalized linear model found a complex relationship between rural living and COPD-related outcomes in US veterans [21]. Thus, a variety of ML models have been successfully applied for use in public health scenarios in general and COPD in particular. The one that ultimately works best in a given situation depends on many factors.

Different races and ethnicities may have different baseline rates of disease due to various factors, including historical misdiagnosis and mistreatment of various racial or ethnic groups, which leads to differential outcomes [22]. There may be outcome, equity, and counseling differences by gender as well as race or ethnicity in the diagnosis and treatment of COPD [23,24].

We had three general expectations of COPD in our models:

1. Cigarette smoking will have the highest impact on COPD rates.
2. AQI will have a strong impact on COPD rates.
3. There will be differences in COPD rates based on racial or ethnic demographics.

Methods

Overview

This paper used MLR and ML methods to predict COPD at the core-based statistical area (CBSA) level [25]. At the time of

this study, there were 388 metropolitan and 541 micropolitan statistical areas in the United States. The data sources were obtained from data repositories of 3 official US agencies, specifically from the CDC. We gathered, integrated, and checked them for data quality. By combining different variables from this variety of data sources, we aimed to obtain a uniquely high accuracy model, while simultaneously reducing biases or flaws that may be attributable to individual data sources. We further checked for missing values (ie, NULL or NA) in every variable. We checked for data correctness by checking the plots of the distributions for every variable, looking for impossible or outlying values. Table 1 shows the data sources used.

Table 1. Data sources.

Source	Reference
National Center for Health Statistics	[26]
Chronic Disease Indicators data	[27]
US Chronic Disease Indicator, stratification values	[28]

Data were collected for all CBSAs that were available from 2016 to 2019. All data obtained from the CDC were contributed voluntarily at the individual level and aggregated to remove all personally identifiable information [29].

The COPD rates are for 2019, whereas all the predictor variables are averaged over the timespan from 2016 to 2018. As such, the models obtained are predictive over time. The data collection result was 517 (56%) of the 929 CBSAs, with proportionally more from the 388 metropolitan statistical areas than from the 541 micropolitan statistical areas. The response variable is the percentage of the CBSA having COPD. We modeled all factors as random variables directly contributing to COPD, which is measured as the proportion (percentage) of the population having COPD. Race or ethnicity was also modeled as percentage of the population rather than as categorical variables. All variables in Table 2 are averaged as mean, except for household income, which was averaged as median.

In Figure 1, we observe that some variables (ie, population, gross domestic product [GDP], GDP per capita, and median household income) are skewed in their distribution.

Table 2. Main variables and descriptive statistics and average within core-based statistical areas.

	Years	Values, median (IQR)	Values, mean (SD)	Values, range
Population (n)	2016-2018	96,811 (48,763-180,484)	191,892 (408,308)	7351-6,633,096
GDP[a] (US $)	2016-2018	13,126,907 (2,562,704-39,046,120)	64,223,036 (212,975,821)	447,355-3,218,209,695
Median household income (US $)	2016-2018	52,632 (46,867-60,494)	54,736 (11,319)	27,842-119,332
GDP per capita (US $)	2016-2018	100.07 (47.83-277.17)	253.77 (479)	16.86-4731.50
Air quality index	2016-2018	38.67 (34.00-43.00)	38.02 (10)	9.00-95.00
Smoking rate	2016-2018	17.12 (15.33-19.28)	17.29 (3)	8.41-29.59
Poverty rate (all ages)	2016-2018	13.80 (10.92-17.12)	14.36 (4)	3.87-35.56
Unemployment rate	2016-2018	4.52 (3.67-5.43)	4.71 (2)	1.97-20.93
Education rate	2016-2018	22.91 (17.94-27.96)	24.22 (8)	8.77-65.75
White (%)	2016-2018	87.6 (78.6-92.8)	84.6 (0.129)	22.1-100
Black (%)	2016-2018	4 (1.5-12.5)	9.3 (0.124)	0.3-100
AI or AN[c] (%)	2016-2018	0.7 (0.4-1.7)	2 (0.044)	0.1-45.9
Asian (%)	2016-2018	1.6 (0.9-3)	2.8 (0.041)	0.2-42.8
NH or PI[d] (%)	2016-2018	0.1 (0.1-0.2)	0.3 (0.01)	0.0-12.9
Hispanic (%)	2016-2018	7 (3.9-14.9)	13.3 (0.164)	0.9-95.5
COPD[b] rate (%)	2019	6.7 (5.7-7.9)	6.871 (1.511)	3.2-15

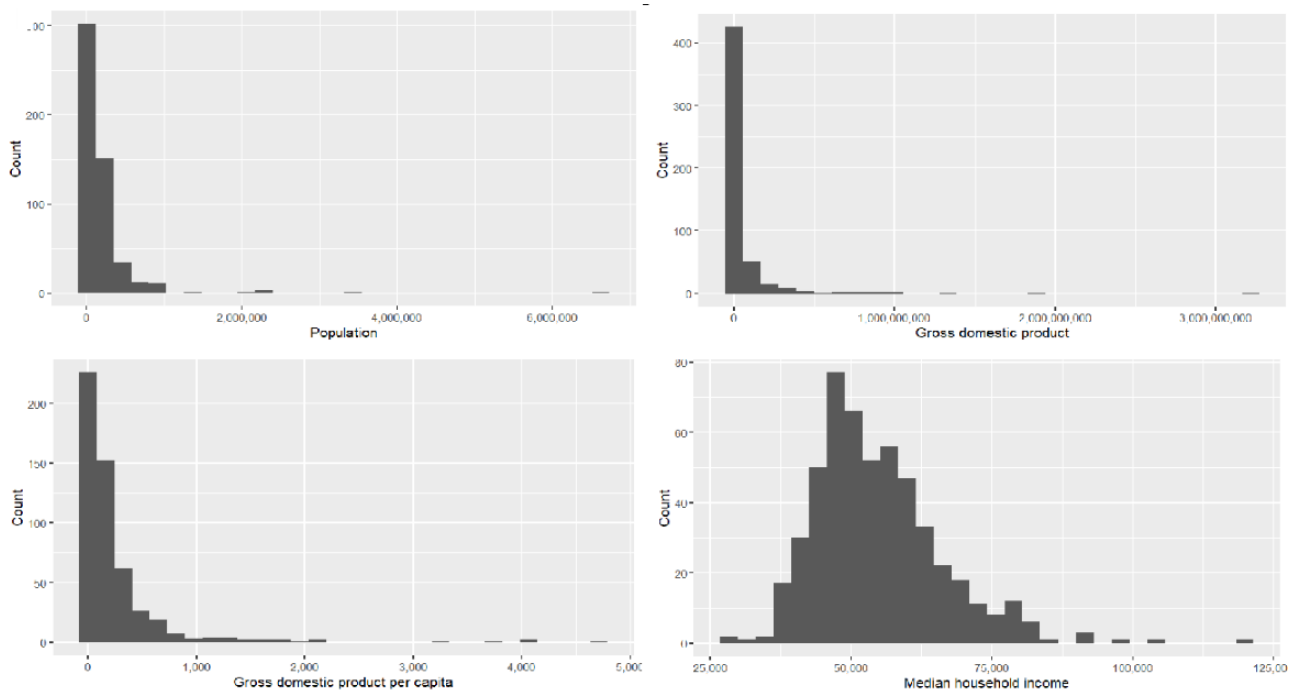
[a]GDP: gross domestic product.

[b]COPD: chronic obstructive pulmonary disease.

[c]AI or AN: American Indian or Alaska Native.

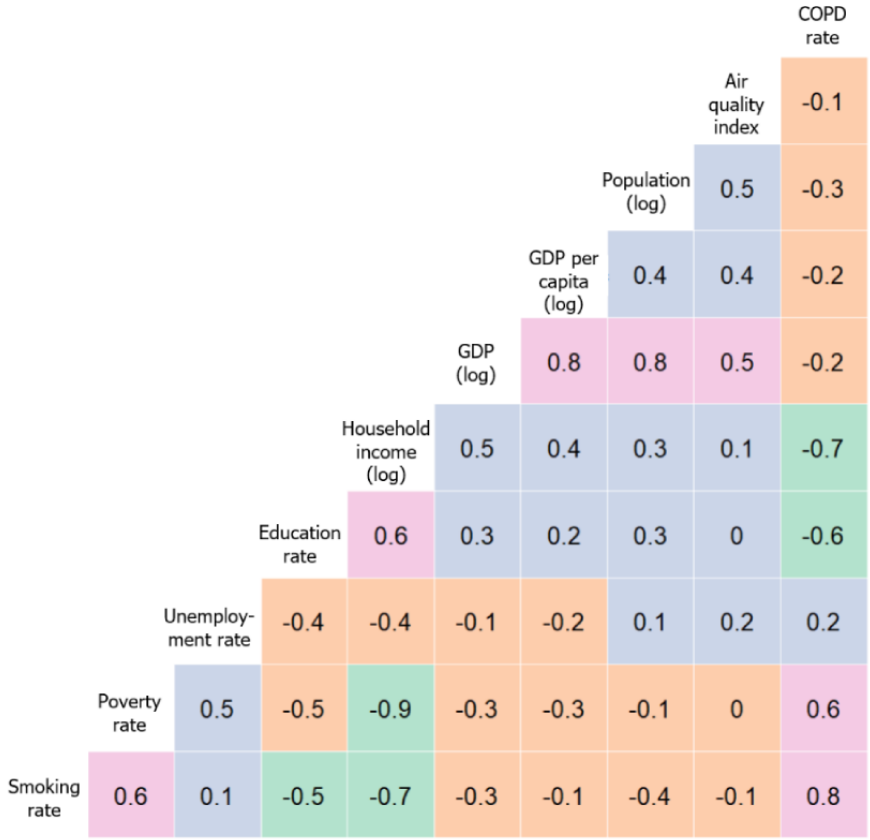
[d]NH or PI: Native Hawaiian or Pacific Islander.

Figure 1. Population, gross domestic product (GDP), GDP per capita, and median household income.



Therefore, we made a log transformation of these variables (ie, logPopl, logGDP, logGDPpc, and logHHI) to make them less skewed, and we show a heat map of correlations of them with the other variables in Figure 2.

Figure 2. Correlations among main variables. COPD: chronic obstructive pulmonary disease; GDP: gross domestic product.



We see a range of correlations, from very negative (green) to negative (orange) to positive (purple) to very positive (pink). In the rightmost column, we see the correlations between the response variable, COPD rate, and the other variables, ranging from very positive (smoking rate) to moderately positive (poverty and unemployment rates) to moderately negative (education and logged household income) to slightly negative (log of GDP, log of population, log of GDP per capita, and

AQI). Given these correlations, we are likely to find good predictive models, but we need to check for multicollinearity in any linear model that we identify.

To understand and model COPD, one has to consider the consistently largest contributing factor: cigarette smoking. Research tends to either control for cigarette smoking or exclude it entirely. In this paper, we chose to include cigarette smoking, accounting for it in our models, but also to examine other factors to compare the magnitudes of influence among the various factors. We aimed to model a variety of factors, including cigarette smoking, to arrive at the model that predicts COPD with the greatest accuracy.

Statistical Analysis

Overview

Our MLR baseline model in R (version 4.2.3) yielded the output in Table 3, which is sorted by absolute value of the *t* value, from high to low.

Table 3. Multiple linear regression.

	Estimate	SE	*t* test (*df*=503)	*P* value
(Intercept)	32.4000	2.930	11.065	<.001
Smoking_Rate	0.2570	0.015	16.635	<.001
Log_HH_Income	−2.8100	0.264	−10.638	<.001
Hispanic_percentage	−2.3900	0.234	−10.249	<.001
Education_Rate	−0.0334	0.005	−6.627	<.001
AI_or_AN_percentage[a]	−2.9000	0.778	−3.726	<.001
Black_percentage	−1.2700	0.356	−3.558	<.001
NH_or_PI_percentage[b]	−15.8000	4.430	−3.558	<.001
Log_GDP	0.0741	0.035	2.126	.034
White_percentage	−0.7380	0.358	−2.060	.04
Unemployment_Rate	0.0456	0.023	1.993	.047
Asian_percentage	2.4800	1.250	1.988	.047
Log_Population	0.0899	0.055	1.626	.105
Air_Quality_Index	0.0003	0.003	0.098	.922

[a]AI_or_AN: American Indian or Alaska Native.

[b]NH_or_PI: Native Hawaiian or Pacific Islander.

The model has residual SE 0.658 on 503 *df*. The multiple R^2 is 0.8152 and adjusted R^2 is 0.8105. The *F*-statistic is 170.7 on 13 and 503 *df* (*P*<.001). The variance inflation factors were checked, with all values <5 indicating low multicollinearity.

There are 7 predictors of high statistical significance: smoking rate, Black percentage, Native Hawaiian or Pacific Islander percentage, American Indian or Alaska Native percentage, education rate, Hispanic percentage, and log of household income. Smoking rate has a positive association with COPD, with every additional percentage increase associated with a 0.257% increase in the COPD rate. The other 6 highly significant predictors have a negative association. Every percentage increase in the log of household income lowers the COPD rate by 2.81%. The Hispanic percentage is nearly as strong; every percentage increase corresponds to a drop of 2.39% in COPD rate. American Indian or Alaska Native is a bit stronger in its coefficient estimate; every percentage point increase corresponds to a drop of 2.9% in COPD rate. Every percentage point increase in Native Hawaiian or Pacific Islanders corresponds to a drop of 15.8%, which is much stronger. Every percentage point increase in Black percentage corresponds to a drop of 1.27% in COPD rate. Education rate has a strongly statistically significant relationship, but a small percentage point impact: every percentage increase corresponds to a decrease of 0.0334% in COPD rate. The remaining 4 predictors—White percentage, GDP (logged), unemployment rate, and Asian percentage—are far less statistically significant and, therefore, should be interpreted with caution.

Linear models are simpler than ML models, and they are sometimes perfectly adequate for explaining a phenomenon. They are easier to interpret, communicate, and implement as new policy. They make statistical assumptions, which can be verified. Linear regression is certainly a good place to start. However, we argue that one should not stop there because an ML model can capture substantial variance from nonlinear relationships (if there are any) in the data and thus produce a more accurate model. By capturing additional variance, the model can capture subtler effects and relationships due to interactions, context, and tipping points. This is crucial because public health practice tends to use simple if-then rules, that is, decision trees. ML models can add nuance to those decision trees based on the captured nonlinearities. Although an adjusted R^2 of 0.8105 looks quite strong, we can perhaps do better with ML methods [18-21].

The 7 ML methods evaluated in this paper are lasso regression, ridge regression, generalized additive model, support vector machine, artificial neural network, random forest, and GBT. These methods were selected for their known strengths in minimizing errors of bias or errors of variance, that is, their ability to fit data well on test data without overfitting. They also represent the range of algorithms commonly used in ML prediction, from methods established in classical statistics to more modern methods derived from computer science. They are commonly used because they are accurate and well understood. Trying a variety of methods is a common practice because the different methods make different statistical assumptions, which may enhance or inhibit optimal performance. All methods were available as R packages for R (version 4.2.3). We summarize each method in terms of its main pros and cons:

Lasso Regression (L1 Regularization)

Lasso regression is an MLR method that incorporates regularization to perform variable selection. It minimizes the sum of squared errors between predicted and actual values, while adding a penalty term based on the absolute value of coefficients multiplied by a tuning parameter. Doing so shrinks some coefficients to exactly 0, effectively performing feature selection by excluding less important variables from the model. This reduces model complexity and minimizes multicollinearity. This is a standard refinement of MLR (R package glmnet).

Ridge Regression (L2 Regularization)

Ridge regression is an MLR technique that adds a penalty term to the objective function to reduce the coefficients of less important predictors and guard against overweighting the most important predictors. While it retains all predictors in the model, ridge regression can help improve the robustness of the model in the presence of correlated predictors by reducing multicollinearity. This is a standard refinement of MLR (R package ridge).

Generalized Additive Model

The generalized additive model is a nonparametric generalization of MLR, which allows for nonlinear terms and coefficient regularization while maintaining interpretability. Each term is a function of X_n rather than simply a numeric coefficient multiplied with X_n. As with MLR, all the terms are added together. Although overfitting can occur, regularization and cross-validation help to minimize it (R package mgcv).

Support Vector Machine

Support vector machine is a technique that transforms the data into a high-dimensional variable space using a kernel function, fitting a function that best fits the data while allowing a certain margin of error (epsilon) and maintaining robustness against outliers. Epsilon tubes can provide a visual representation of the model's uncertainty. Points within the tube are considered well predicted, while those outside represent errors. A regularization parameter controls the trade-off between accuracy and complexity (R package e1071).

Artificial Neural Network

Artificial neural network is a generalization of MLR with hidden layers of nodes between input and output nodes; it may result in overfitting. Depending on the number of hidden layers, nodes per layer, and the activation function used to convert inputs to outputs, an arbitrarily complex model can be fit. This can be thought of as a simplified version of a human brain, in which input and output nodes are separated by ≥1 layers of hidden nodes. Prediction error causes the weights of the hidden nodes to be adjusted until minimal error is achieved (R package neuralnet).

Random Forest

Random forest is an ensemble technique to fit a large number of a bootstrap-sampled aggregation (bagging) of trees by considering a random subset of variables at each tree split. Intuitively, a random forest is a blending of a large number of decision trees, the "wisdom of the forest." The random subset of variables restriction is done to prevent strong variables from dominating the weaker variables. A random forest tends to perform very well but is difficult to interpret (R package RandomForest).

Gradient Boosted Trees

GBT is an ensemble of sequential trees that focuses on the errors of the previous tree. It is able to find interaction effects implicitly. It uses gradient descent search to rapidly minimize error via an arbitrary, differentiable loss function. It uses many trees to help ensure that the local minimum error found is the global minimum. Intuitively, this builds a strong predictive model by combining many weak models, each correcting the errors of the previous one (R package XGBoost).

Our ML approach followed best practices. We randomly partitioned the data set into train (311/517, 60%), cross-validate (103/517, 20%), and test (103/517, 20%) subsets. We checked for outliers, multicollinearity, and target leakage to ensure valid models [30].

Ethical Considerations

This research did not involve human subjects at the individual level and therefore did not require institutional review board approval. Our data were collected from CDC sources at the level of CBSA. All sources were free of personal identifying information, because the CDC is legally required to ensure the protection of the data. All data were collected and aggregated in a non–personal identifying information manner. The results of our analysis do suggest communicating with different racial and ethnic groups differently, tailoring the implications directly to patients as well as indirectly to their families, communities, and health care providers in a race- or ethnicity-sensitive manner.

Results

In Table 4, we describe the results of the ML models of COPD by various accuracy metrics. For the accuracy metrics, we used 3 standard measures of predictive accuracy in addition to variance explained (adjusted R^2): root mean square error (RMSE), mean absolute error (MAE), and symmetric mean

absolute percentage error (SMAPE) [31,32]. We performed a grid search over all the main numeric parameters for a given method to find the optimal combination of parameter values [33]. A grid search tries all combinations of parameters from a minimum to a maximum value by some step size. Those minimum, maximum, and step sizes are determined from typical default values and best practices. The best metrics in Table 4 are indicated by italics.

Table 4. Machine learning models versus multiple linear regression.

Method	Adjusted R^2	Root mean square error			Mean absolute error			Symmetric mean absolute percentage error		
		Train	CV[a]	Test	Train	CV	Test	Train	CV	Test
Gradient boosted tree (XGBoost, loss function=least squares, learning rate=0.05, and maximum tree depth=10)	0.857	0.550	0.598	0.557	0.433	0.445	*0.456* [b]	6.473	6.543	*6.956*
Support vector machine (Nystroem kernel and loss function=Poisson deviance)	*0.858*	0.555	*0.558*	*0.556*	0.435	*0.434*	0.462	6.515	*6.443*	6.989
Random forest (maximum trees=500, maximum depth=none, and maximum leaves=100)	0.836	*0.534*	0.614	0.596	*0.420*	0.462	0.479	*6.315*	6.819	7.339
Neural network (2 layers: 512, 512 units; regularization via random dropout rate=0.05 and activation function=prelu)	0.845	0.601	0.609	0.580	0.455	0.467	0.468	6.856	6.928	7.182
Generalized additive model (learning rate=0.3, maximum bins=100, and loss function=least squares)	0.822	0.629	0.658	0.621	0.515	0.508	0.488	7.619	7.502	7.212
Ridge regression	0.810	0.589	0.618	0.641	0.467	0.483	0.527	6.986	7.346	7.986
Lasso regression	0.758	0.750	0.778	0.724	0.585	0.593	0.597	8.425	8.544	8.824
Multiple linear regression	0.811	0.620	0.699	0.749	0.474	0.548	0.591	7.205	8.403	9.666

[a]CV: cross-validation.

[b]Values in italics represent the best metrics.

The ML methods were superior to MLR on most metrics. Support vector machine was the best on adjusted R^2 and RMSE, slightly superior to GBT, but GBT was superior by a larger margin on MAE and SMAPE. Therefore, we chose GBT as the best overall method. In Multimedia Appendix 1, we show the variable importance plot for the GBT model. Variable importance plots are a common first way to peer inside a "black-box method" and understand the relative importance of the variables used within it [34].

The top five variables in terms of impact were (1) smoking rate and (2) household income, followed by (3) American Indian or Alaska Native percentage, (4) education rate, and (5) unemployment rate. Black percentage was sixth, Hispanic percentage was seventh, and there was only a small impact from the remaining variables: White percentage, AQI, Asian percentage, Native Hawaiian or Pacific Islander percentage, population, and GDP. Relative to the MLR, smoking rate, household income, education rate, and Black percentage remained the same in terms of rank importance. Hispanic percentage dropped from third to seventh rank; American Indian or Alaska Native percentage rose from fifth to third rank; and unemployment rate rose sharply, from 10th to 5th in importance. Native Hawaiian or Pacific Islander percentage dropped sharply, from 7th to 11th in rank.

Figure 3 shows the lift plot, and Figure 4 shows the predictive residual plot. The lift plot shows observations sorted by predicted value deciles. The ratio of the observed outcome to the expected outcome was calculated and plotted. The predictive residual plot shows the differences between observed and predicted values.

Figure 3. Lift plot showing chronic obstructive pulmonary disease rate as a function of 10 decile bins; predicted values are in blue and actual values are in red. COPD: chronic obstructive pulmonary disease.

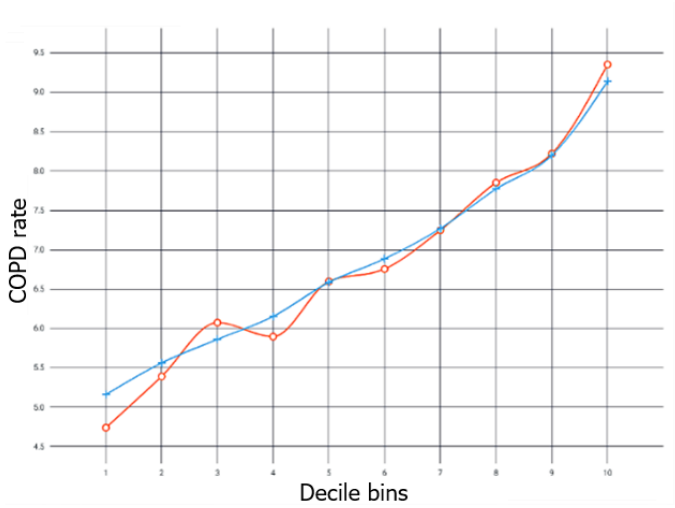
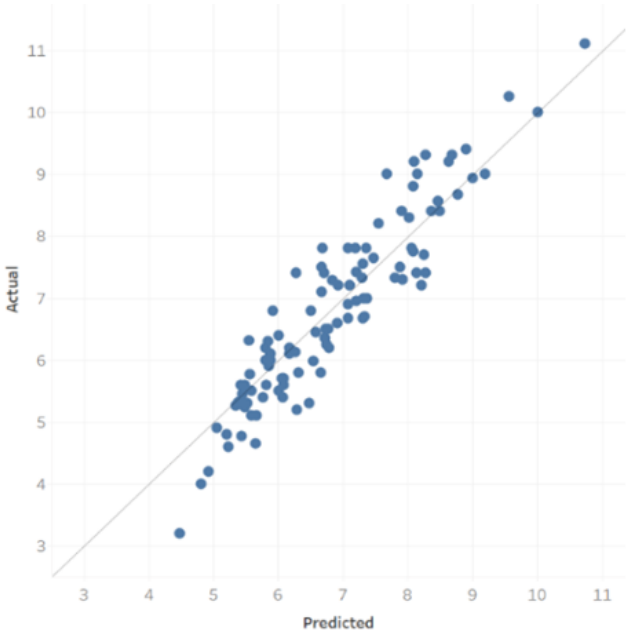


Figure 4. Prediction residuals.



In addition to the variable importance plot, other plots were used to gain an understanding of ML models: local interpretable model-agnostic explanations (LIME) models and SHAP (Shapley additive explanations) plots [35-37]. We chose SHAP plots because they are based on a cooperative game-theoretical foundation, showing every combination of the variables in the model and how they work together to predict the outcome variable. Figure 5 shows the SHAP plot for all the GBT's variables.

The top 5 variables (smoking rate, household income, American Indian or Alaska Native percentage, education rate, and unemployment rate) have substantially more impact on COPD percentage than the remaining variables. We show the top 5 variables as well as the next 4 as individual SHAP plots of the GBT in Figure 6. All 9 plots show significant nonlinearities.

Smoking had the greatest impact: as the smoking rate increased, the COPD rate also rose substantially, following a steeply curved, nearly exponential relationship. Median household income had the second highest impact, an almost linear (and negative) relationship. The greater the household income, the lower the COPD rate. This could indicate better insurance coverage, better health care access, higher quality health care (ie, prevention or treatment), lower occupational exposure, or lower home exposure (eg, gas stoves). The next variable was American Indian or Alaska Native percentage, indicating a negative but nonlinear relationship with COPD rate: a steep drop followed by a gradual tapering. This represents a significant protective influence shown for the American Indian or Alaska Native community, which has not yet been noted in the literature.

Figure 5. SHAP (Shapley additive explanations) values for all features (variables). AI: American Indian; AN: Alaska Native; GDP: gross domestic product; NH: Native Hawaiian; PI: Pacific Islander.

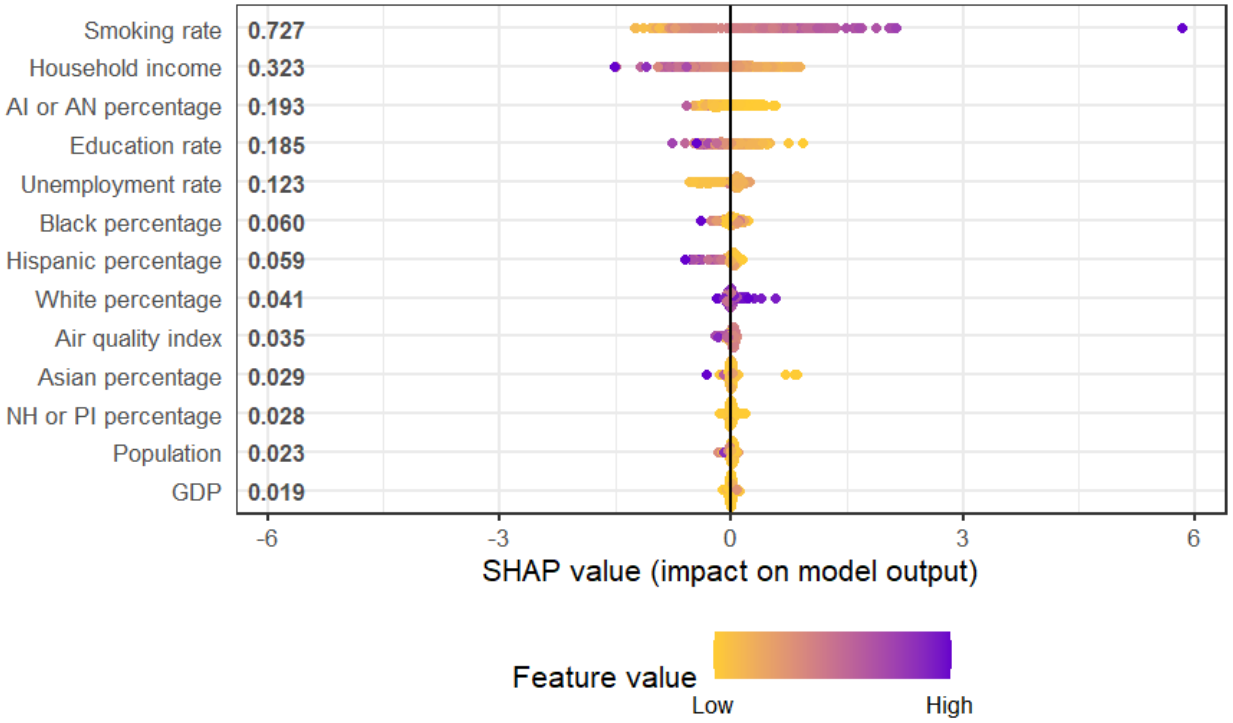
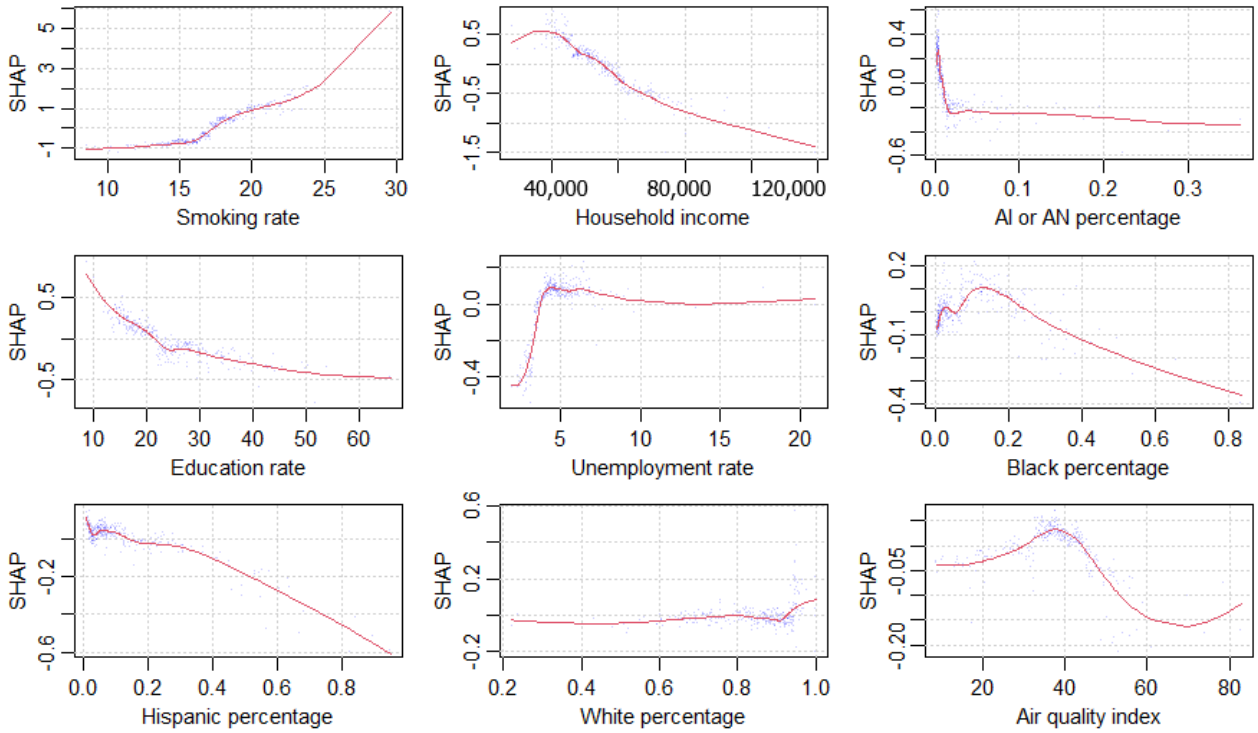


Figure 6. SHAP (Shapley additive explanations) plots for the 9 most important variables. AI: American Indian; AN: Alaska Native.



The next variable, education rate had a negative, curvilinear relationship. The more educated the population, the lower the COPD rate. The explanation could be similar to that of income: better insurance coverage or health care access, better quality of health care, lower occupational exposure, or lower home exposure [38]. The next variable was unemployment rate, with a sharply positive but flat relationship with COPD rate. The next variable was Black percentage, with an initial positive relationship with COPD rate but then a reversal to a negative, linear relationship.

The next variable, Hispanic percentage, showed a negative linear relationship with COPD rate. This represents a significant protective influence shown for people in the Hispanic

community, which is consistent with the literature [39-45]. The next variable was White percentage, showing a slightly negative relationship with COPD rate. Finally, the last variable was AQI (higher value being worse), which shows an initial positive relationship with COPD rate, peaking around 38. This may be a critical point, after which people take precautions not to be exposed to the low-quality air.

Discussion

Principal Findings

We had three general expectations, which were largely met:

1. The impact of cigarette smoking was the largest in all models.
2. The AQI had an impact in the best ML model, but it was smaller than expected.
3. There were substantial racial or ethnic differences, particularly among American Indian or Alaska Native, Black, and Hispanic communities.

Consistent with the literature, we found that smoking remains the most significant risk factor for COPD, with research consistently demonstrating a strong association between smoking status and COPD prevalence. In our MLR, we found that smoking rate is the strongest predictor of COPD rate. We found the same result in our GBT but also found that the smoking rate has a curvilinear, almost exponential, relationship with COPD. The Rotterdam study, a large-scale population-based cohort study, found that current and former smokers had a substantially higher risk of developing COPD compared to never smokers [46]. A nationwide population-based cohort study in South Korea demonstrated that smoking cessation after COPD diagnosis was associated with lower all-cause and cause-specific mortality [47].

Notably, 3 of the 4 next most important variables, in terms of impact in our GBT, are socioeconomic variables: household income (rank 2), education rate (rank 4), and unemployment rate (rank 5). In the MLR, we found that household income (logged) had the second highest impact. In the GBT, household income had the second highest impact, but the tipping point was around US $40,000, after which higher income had a linear, negative relationship with COPD. Education rate had a strongly negative, curvilinear relationship with COPD. Unemployment rate had a sharply positive relationship with COPD, but then peaked at 5% unemployment, after which it plateaued.

These results are largely consistent with the literature on socioeconomic factors and smoking behavior, suggesting an indirect relationship with COPD via smoking. A study examining smoking among adolescents in 6 European cities found that disposable income was positively associated with smoking [48]. Conversely, lower socioeconomic status was associated with higher COPD prevalence because in addition to lower education and income, there may be environmental pollutants, occupational hazards, or barriers to COPD screening, diagnosis, and treatment [49]. In contrast with the literature, our SHAP plots show mostly nonlinear relationships with COPD. Household income showed a tipping point at US $40,000, after which the negative relationship with COPD was nearly linear.

Ethnic or racial variables accounted for 3 of the top 7 variables in the GBT: American Indian or Alaska Native percentage (rank 3), Black percentage (rank 6), and Hispanic percentage (rank 7). The greater the size of those minority populations, the lower the COPD rate. Our SHAP plots show significant tipping points (nonlinearities) for American Indian or Alaska Native percentage and Black percentage and a mostly linear relationship for Hispanic percentage. Consistent with the literature, all 3 variables show a strongly negative association with COPD.

The regression and GBT models show that in addition to strongly protective impacts for lower cigarette smoking and higher household income, there are protective impacts for larger American Indian or Alaska Native and Hispanic populations as well as a nonlinear impact on larger Black populations. Higher education rate and lower unemployment rate are also protective, whereas AQI shows mixed effects. These results have implications for private health care practitioners, public health care officials, and health care policy makers who aim to reduce COPD rates. Such policies and programs should not assume high digital literacy [50,51]. System designers could use SMS text messaging, social media, and interactive voice response systems. This would be appropriate for those with lower household income or lower education levels. To design culturally appropriate visual cues and messaging to different racial or ethnic groups, members of the various communities should be included in the design process [52,53]. In sum, the user interface should exhibit high ease-of-use—using gamification, storytelling, and peer support—consistent with cultural norms.

Several studies have identified ethnic and racial disparities in COPD prevalence and risk among smokers. One study found that racial and ethnic minority individuals, particularly African Americans and Hispanics, had a lower prevalence of airflow obstruction than non-Hispanic White individuals, even after adjusting for smoking status and other risk factors [54]. This finding was supported by another study that observed lower COPD risk in ethnic minority groups compared to White individuals, despite similar smoking intensities [55]. A larger minority population means a larger peer support network for prevention and cessation of smoking and a larger peer community to recommend COPD screening, diagnosis, and treatment, which is particularly useful in a health care system that has implicit racial or ethnic bias [50,56].

There are varying levels of patient trust and implicit bias in health care practitioners themselves [57], which contributes to health outcome differences. From a population communication perspective, messaging regarding the risks of COPD—particularly the avoidance or cessation of cigarette smoking—should be sensitive to community context, engaging trusted local authorities to optimize the chances of patient engagement [58]. Health care practitioners could partner with trusted local authorities and community leaders regarding smoking prevention and cessation as well as respiratory health in general to decrease COPD risk. Health care practitioners and

educators should communicate to different populations in culturally sensitive ways [59,60].

Educational materials and behavior change strategies may need to be customized according to different risk factors, beliefs, preferences, and technographics of different subpopulations [50,51]. On a basic level, people with lower levels of education or household income could be directed via phone geolocation to their local health care and to their community leaders for in-person guidance and support. Local leaders could then inform them about local smoking cessation programs and apps or websites that monitor air quality in their community. Trusted local authorities are helpful entry points in those communities, after which peer support and network effects spread the information.

AQI was not significant in the MLR, but it was significant in the GBT, albeit not as strongly as we expected. It could be that the AQI is more of a diffuse, macrolevel environmental factor that fluctuates over time, making some CBSAs worse on average, but with wide volatility, for example, as weather and wind directions change [61,62]. Therefore, AQI could have more of an indirect or interaction effect with other variables. Combining campaigns on smoking prevention with campaigns on air quality could create a holistic public health strategy, particularly—as our findings suggest—in communities considered vulnerable, that is, communities with lower education, higher unemployment, and lower household income. Subsidies for households in communities considered vulnerable to convert to more efficient, cleaner home heating and cooling methods would improve their home's air quality at a lower cost [63]. Research suggests that engaging communities in targeting their air quality issues can lead to more positive outcomes in both air quality and public health [64-66].

There is a small but growing body of research that uses ML models in health care and medicine. There is recognition that the models can be highly accurate, but there is no consensus yet on how to interpret the results in a way that meshes seamlessly with clinical practice. The following examples provide an overview.

Elshawi et al [67] compared model-agnostic explanations using 2 techniques, LIME and Shapley values, to interpret a ML model for predicting hypertension risk. LIME uses small subsets of the data, which may be idiosyncratic, to provide intuitive explanations, that is, rules. Shapley values are more theoretically sound and global, using all the available data, and are, therefore, less idiosyncratic than LIME, but they do not provide LIME's simple, linear explanations [67].

Hakkoum et al [68] conducted an extensive literature review of ML interpretability in medicine published between 1994 and 2020. The review found that there was no consensus on evaluation metrics or frameworks to assess the quality and utility of the interpretability methods [68]. The highest performing ML models did not translate easily into clinical rules.

Meng et al [69] reviewed the interpretability and fairness evaluation of deep learning models on MIMIC-IV data set, a large, publicly available benchmark for developing and evaluating the interpretability of high-performing ML models that use sensitive demographic features. The review found that existing interpretation methods, for example, variable importance rankings, provide partial explanations without fully elucidating the model's complex decision logic.

In sum, there is no consensus on the best way to interpret high-performing ML models in health care. There are always trade-offs between accuracy and interpretability or explainability. We chose to use Shapley values because they represent the frontier in explainability, and they are similar to interpreting a multiple regression, interpreting 1 variable at a time, without the assumptions of linear models. In addition, Shapley values allow for nonlinear relationships between each independent (predictor) variable and the dependent variable. Variable importance plots in conjunction with Shapley values help us to identify the most important variables and characterize their relationships with COPD.

Our best MLR model had variance explained of 81.1%, MAE of 0.591, and SMAPE of 9.666. Our best ML model was the GBT, with variance explained of 85.7%, MAE of 0.456, and SMAPE of 6.956. The GBT explains most of the variance—4.6% more than the best MLR—with far less predictive error. The GBT's SMAPE (6.956) was 28% lower than that of the MLR's SMAPE (9.666). Similarly, the GBT's RMSE was 26% lower than the MLR's RMSE, and its MAE was 23% lower than that of the MLR. Real-world predictive accuracy should be similar to that found in the test data set because the test data were never used in the GBT's model development.

Our GBT performed strongly on the test data, with very little performance deterioration on the test data versus performance on the training and validation data. This demonstrates that the GBT model does not overfit the data. To interpret the GBT, we used a variable importance plot [34,70,71] and SHAP plots [72,73]. SHAP plots are useful for interpreting the strength of the pairwise relationships between predictor variable and COPD rate, showing the added nuances of the curvilinear plots. By doing so, we rendered transparent the "black-box model" [74-76], thus preserving interpretability and actionability, in addition to adding nonlinear nuance.

Limitations and Future Directions

This research has a few limitations. The data were obtained from 517 (56%) of the 929 CBSAs. We assumed that this was an adequate sample and that the remaining CBSAs that did not report the data were similar to those that did. Alternatively, it could be that the CBSA that did not report COPD rates did so because the rates were low, that is, COPD was not considered a major problem by the local public health officials. Data covering additional demographic variables, such as gender and age, in addition to occupational exposures and physical exercise, could be gathered [77-79]. Future research could develop separate models stratified by demographic variables such as race or ethnicity, assuming there are sufficient data for each categorical class. There could also be geopolitical variations in terms of population density as well as demographics, psychographics [80], and technographics [81,82].

Future data collection could focus on understanding racial or ethnic disparities. By collecting data more intensively from the minority populations, we could go deeper into understanding how their rates of COPD drop so dramatically. Is it related to active peer recommendations for better self-care in a predominantly White health care system and population? Is it related to successfully tailored smoking prevention or cessation programs? Data pertaining to answering these more specific questions could be collected to enhance our understanding of how best to tailor communications to different demographic or ethnographic groups.

All our models were structured as direct effects. We applied MLR and ML methods with data from CBSAs, which have significant variation in terms of health care access and quality. Using these models as a foundation, we should recognize the interconnectedness (ie, direct, indirect, and interactive) of pollutants and copollutants to fully understand COPD's complex etiology. Future research could model interaction, moderating, or mediating effects, perhaps with a structural equation model, to identify the direct and indirect effects of COPD, for example, showing how asthma may lead to COPD or to asthma-COPD overlap [77].

There are many research knowledge gaps in the health impacts of extreme air pollution, including the effects of interactions between temperature and air pollution on respiratory health due to climate change [83]. Future research directions could focus on modeling the direct and indirect links between environmental exposures and COPD. On the basis of those results, we could design interventions, such as air quality warning systems, to mitigate their impact. The findings would underscore the opportunities for public health regulations, public-private sector partnerships, private company entrepreneurship, and global initiatives to reduce environmental exposures.

Greenhouse gas emissions may exacerbate overall air quality [84-88], contributing indirectly to COPD. Future research could collect data on new, additional variables pertaining to climate change [89]. Wildfires, which are increasingly common, produce more carcinogens in the air, including high levels of particulate matter. This can directly decrease air quality or copollute with other ambient pollutants [90]. These problems have been shown to increase the odds of lung cancer [91], and it is plausible that they can also contribute to COPD.

The association between COPD and environmental pollutants, including tropospheric ozone, nitrogen dioxide, sulfur dioxide, and occupational exposures, has been extensively investigated [8,91-94]. Coarse, fine, and ultrafine particulate matter have been studied extensively and linked to systemic oxidative stress, inflammation [95], atherosclerosis [96], and mortality [97] in the United States [98,99] and China [100-102]. Tropospheric ozone exposure by itself has been linked to impaired lung function and increased COPD-related hospital admissions [103-105]. Similarly, elevated levels of nitrogen dioxide and sulfur dioxide, which are common in cities and industrial work sites, have been linked to an increased risk of COPD in the general population [106,107] and older adults [108]. In sum, data pertaining to ambient pollution, for example, particulate matter, sulfur dioxide, and carbon monoxide, could be useful additional copollutant data to include in future models [6,86-88,91,109-111].

Conclusions

Our novel contributions in this paper include the following: (1) integration of multiple publicly available CDC data sources, (2) development of highly accurate models using linear and nonlinear methods, and (3) interpretation of the variable impacts for the best model. Smoking was the number 1 variable impacting the COPD rate, which was expected. Household income was the second most influential predictor variable. Four economic factors spanned the full range of influence, from large (household income) to moderate (education rate) to small (unemployment rate and GDP). The race or ethnicity variable also had a range of impacts, from moderately high (American Indian or Alaska Native percentage) to moderate (Black or Hispanic percentage) to small (White, Asian, or Native Hawaiian or Pacific Islander percentage).

This research demonstrates the power of ML methods in general and a GBT, which produced a highly accurate model of COPD rates. The computational complexity of a GBT enables it to obtain high accuracy, but health care policy makers may be reluctant to adopt it unless they can obtain a rule-based explanation. Furthermore, clinicians typically want to be able to explain, justify, and communicate results to others in an intuitive manner. Finally, there may be legal, auditing, or regulatory requirements concerning transparency. If the method is audited, and it cannot be clearly explained, there may be serious legal or financial consequences [72]. Consequently, it is important to have explainable models to open the "black box," rendering them interpretable and actionable [75,76]. This research shows that it is possible to do so.

Acknowledgments

The authors thank Malavika Andavilli, Xuhui Bai, Chulin Chen, Shan He, Lanxiang Shao, Moshe Shpits, and Ziyu Tang for contributions to an early version of this paper. The authors also thank the anonymous reviewers at the International Business School, Brandeis University, for their helpful comments.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Importance of model variables for gradient boosted tree.

[PNG File , 10 KB-Multimedia Appendix 1]

References

1. Diaz-Guzman E, Khosravi M, Mannino DM. Asthma, chronic obstructive pulmonary disease, and mortality in the U.S. population. COPD. Dec 08, 2011;8(6):400-407. [doi: 10.3109/15412555.2011.611200] [Medline: 22149399]
2. Hardin M, Silverman EK, Barr RG, Hansel NN, Schroeder JD, Make BJ, et al. COPDGene Investigators. The clinical features of the overlap between COPD and asthma. Respir Res. Sep 27, 2011;12(1):127. [FREE Full text] [doi: 10.1186/1465-9921-12-127] [Medline: 21951550]
3. Peden DB. Air pollution in asthma: effect of pollutants on airway inflammation. Ann Allergy Asthma Immunol. Dec 2001;87(6 Suppl 3):12-17. [doi: 10.1016/s1081-1206(10)62334-4] [Medline: 11770676]
4. Wong-Parodi G, Dias MB, Taylor M. Effect of using an indoor air quality sensor on perceptions of and behaviors toward air pollution (Pittsburgh empowerment library study): online survey and interviews. JMIR Mhealth Uhealth. Mar 08, 2018;6(3):e48. [FREE Full text] [doi: 10.2196/mhealth.8273] [Medline: 29519779]
5. Iribarren SJ, Akande TO, Kamp KJ, Barry D, Kader YG, Suelzer E. Effectiveness of mobile apps to promote health and manage disease: systematic review and meta-analysis of randomized controlled trials. JMIR Mhealth Uhealth. Jan 11, 2021;9(1):e21563. [FREE Full text] [doi: 10.2196/21563] [Medline: 33427672]
6. Valavanidis A, Vlachogianni T, Fiotakis K. Tobacco smoke: involvement of reactive oxygen species and stable free radicals in mechanisms of oxidative damage, carcinogenesis and synergistic effects with other respirable particles. Int J Environ Res Public Health. Feb 2009;6(2):445-462. [FREE Full text] [doi: 10.3390/ijerph6020445] [Medline: 19440393]
7. Berend N. Contribution of air pollution to COPD and small airway dysfunction. Respirology. Feb 27, 2016;21(2):237-244. [FREE Full text] [doi: 10.1111/resp.12644] [Medline: 26412571]
8. Lissåker CT, Talbott EO, Kan H, Xu X. Status and determinants of individual actions to reduce health impacts of air pollution in US adults. Arch Environ Occup Health. Dec 02, 2016;71(1):43-48. [doi: 10.1080/19338244.2014.988673] [Medline: 25454076]
9. Dransfield MT, Bailey WC. COPD: racial disparities in susceptibility, treatment, and outcomes. Clin Chest Med. Sep 2006;27(3):463-71, vii. [doi: 10.1016/j.ccm.2006.04.005] [Medline: 16880056]
10. Duru OK, Harawa NT, Kermah D, Norris KC. Allostatic load burden and racial disparities in mortality. J Natl Med Assoc. Jan 2012;104(1-2):89-95. [FREE Full text] [doi: 10.1016/s0027-9684(15)30120-6] [Medline: 22708252]
11. Alexander GR, Wingate MS, Bader D, Kogan MD. The increasing racial disparity in infant mortality rates: composition and contributors to recent US trends. Am J Obstet Gynecol. Jan 2008;198(1):51.e1-51.e9. [doi: 10.1016/j.ajog.2007.06.006] [Medline: 17870043]
12. Woolf SH, Johnson RE, Fryer GE, Rust G, Satcher D. The health impact of resolving racial disparities: an analysis of US mortality data. Am J Public Health. Dec 2004;94(12):2078-2081. [doi: 10.2105/ajph.94.12.2078] [Medline: 15569956]
13. Budd J, Miller BS, Manning EM, Lampos V, Zhuang M, Edelstein M, et al. Digital technologies in the public-health response to COVID-19. Nat Med. Aug 07, 2020;26(8):1183-1192. [doi: 10.1038/s41591-020-1011-4] [Medline: 32770165]
14. Zicari RV, Brusseau J, Blomberg SN, Christensen HC, Coffee M, Ganapini MB, et al. On assessing trustworthy AI in healthcare. Machine learning as a supportive tool to recognize cardiac arrest in emergency calls. Front Hum Dyn. Jul 8, 2021;3:7. [doi: 10.3389/fhumd.2021.673104]
15. Wenstrup J, Havtorn JD, Borgholt L, Blomberg SN, Maaloe L, Sayre MR, et al. A retrospective study on machine learning-assisted stroke recognition for medical helpline calls. NPJ Digit Med. Dec 19, 2023;6(1):235. [FREE Full text] [doi: 10.1038/s41746-023-00980-y] [Medline: 38114611]
16. Wu CT, Li GH, Huang CT, Cheng YC, Chen CH, Chien JY, et al. Acute exacerbation of a chronic obstructive pulmonary disease prediction system using wearable device data, machine learning, and deep learning: development and cohort study. JMIR Mhealth Uhealth. May 06, 2021;9(5):e22591. [FREE Full text] [doi: 10.2196/22591] [Medline: 33955840]
17. Peng J, Chen C, Zhou M, Xie X, Zhou Y, Luo CH. A machine-learning approach to forecast aggravation risk in patients with acute exacerbation of chronic obstructive pulmonary disease with clinical indicators. Sci Rep. Feb 20, 2020;10(1):3118. [FREE Full text] [doi: 10.1038/s41598-020-60042-1] [Medline: 32080330]
18. Moll M, Qiao D, Regan EA, Hunninghake GM, Make BJ, Tal-Singer R, et al. Machine learning and prediction of all-cause mortality in COPD. Chest. Sep 2020;158(3):952-964. [FREE Full text] [doi: 10.1016/j.chest.2020.02.079] [Medline: 32353417]
19. Lee Y, Kim E, Chae KJ, Lee CH. Machine learning predicts computed tomography (CT)-based normal regional lung function distribution in asthma and chronic obstructive pulmonary disease (COPD) (abstract). In: Proceedings of the American Thoracic Society 2023 International Conference. 2023. Presented at: ATSIC '23; May 19-24, 2023; Washington, DC. URL: https://tinyurl.com/2fwtcz4b [doi: 10.1164/ajrccm-conference.2023.207.1_MeetingAbstracts.A4004]
20. Estépar RS. Artificial intelligence in COPD: new venues to study a complex disease. Barc Respir Netw Rev. 2020;6(2):144-160. [FREE Full text] [doi: 10.23866/BRNRev:2019-0014] [Medline: 33521399]
21. Fortis S, Gao Y, Baldomero AK, Sarrazin MV, Kaboli PJ. Association of rural living with COPD-related hospitalizations and deaths in US veterans. Sci Rep. May 16, 2023;13(1):7887. [FREE Full text] [doi: 10.1038/s41598-023-34865-7] [Medline: 37193770]

22. Freimuth VS, Quinn SC, Thomas SB, Cole G, Zook E, Duncan T. African Americans' views on research and the Tuskegee syphilis study. Soc Sci Med. Mar 2001;52(5):797-808. [FREE Full text] [doi: 10.1016/s0277-9536(00)00178-7] [Medline: 11218181]
23. Vozoris NT, Stanbrook MB. Smoking prevalence, behaviours, and cessation among individuals with COPD or asthma. Respir Med. Mar 2011;105(3):477-484. [FREE Full text] [doi: 10.1016/j.rmed.2010.08.011] [Medline: 20850288]
24. Ezzati M, Friedman AB, Kulkarni SC, Murray CJ. The reversal of fortunes: trends in county mortality and cross-county mortality disparities in the United States. PLoS Med. Apr 22, 2008;5(4):e66. [FREE Full text] [doi: 10.1371/journal.pmed.0050066] [Medline: 18433290]
25. Housing patterns and core-based statistical areas. United States Census Bureau. URL: https://www.census.gov/topics/housing/housing-patterns/about/core-based-statistical-areas.html [accessed 2024-04-29]
26. NCSH provides timely and accurate health statistics for the United States. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/nchs/ [accessed 2024-08-05]
27. COVID-19 public data set. Centers for Disease Control and Prevention. URL: https://chronicdata.cdc.gov/ [accessed 2024-08-05]
28. Chronic disease indicators (CDI). Centers for Disease Control and Prevention. URL: https://www.cdc.gov/cdi/ [accessed 2024-08-06]
29. Office of public health data, surveillance, and technology (OPHDST). Centers for Disease Control and Prevention. Jan 2024. URL: https://www.cdc.gov/about/divisions-offices/ophdst.html [accessed 2024-04-29]
30. Kaufman S, Rosset S, Perlich C, Stitelman O. Leakage in data mining: formulation, detection, and avoidance. ACM Trans Knowl Discov Data. Dec 18, 2012;6(4):1-21. [doi: 10.1145/2382577.2382579]
31. Hyndman RJ, Koehler AB. Another look at measures of forecast accuracy. Int J Forecast. Oct 2006;22(4):679-688. [doi: 10.1016/j.ijforecast.2006.03.001]
32. Willmott CJ, Matsuura K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim Res. 2005;30:79-82. [doi: 10.3354/cr030079]
33. Fayed HA, Atiya AF. Speed up grid-search for parameter selection of support vector machines. Appl Soft Comput. Jul 2019;80:202-210. [doi: 10.1016/j.asoc.2019.03.037]
34. Greenwell BM, Boehmke BC. Variable importance plots—an introduction to the VIP package. R J. 2020;12(1):343-366. [doi: 10.32614/RJ-2020-013]
35. Antwarg L, Miller RM, Shapira B, Rokach L. Explaining anomalies detected by autoencoders using Shapley additive explanations. Expert Syst Appl. Dec 2021;186:115736. [FREE Full text] [doi: 10.1016/j.eswa.2021.115736]
36. Garreau D, von Luxburg U. Explaining the explainer: a first theoretical analysis of LIME. arXiv. Preprint posted online January 10, 2020.. [FREE Full text] [doi: 10.1016/b978-0-32-396098-4.00020-x]
37. Gramegna A, Giudici P. SHAP and LIME: an evaluation of discriminative power in credit risk. Front Artif Intell. 2021;4:752558. [FREE Full text] [doi: 10.3389/frai.2021.752558] [Medline: 34604738]
38. Hetlevik Ø, Melbye H, Gjesdal S. GP utilisation by education level among adults with COPD or asthma: a cross-sectional register-based study. NPJ Prim Care Respir Med. Jun 09, 2016;26:16027. [FREE Full text] [doi: 10.1038/npjpcrm.2016.27] [Medline: 27279354]
39. DiBonaventura MD, Paulose-Ram R, Su J, McDonald M, Zou KH, Wagner JS, et al. The impact of COPD on quality of life, productivity loss, and resource use among the elderly United States workforce. COPD. Feb 2012;9(1):46-57. [doi: 10.3109/15412555.2011.634863] [Medline: 22292597]
40. Ford ES, Croft JB, Mannino DM, Wheaton AG, Zhang X, Giles WH. COPD surveillance--United States, 1999-2011. Chest. Jul 2013;144(1):284-305. [FREE Full text] [doi: 10.1378/chest.13-0809] [Medline: 23619732]
41. Tilert T, Dillon C, Paulose-Ram R, Hnizdo E, Doney B. Estimating the U.S. prevalence of chronic obstructive pulmonary disease using pre- and post-bronchodilator spirometry: the National Health and Nutrition Examination Survey (NHANES) 2007-2010. Respir Res. Oct 09, 2013;14(1):103. [FREE Full text] [doi: 10.1186/1465-9921-14-103] [Medline: 24107140]
42. Zhang X, Holt JB, Lu H, Wheaton AG, Ford ES, Greenlund KJ, et al. Multilevel regression and poststratification for small-area estimation of population health outcomes: a case study of chronic obstructive pulmonary disease prevalence using the behavioral risk factor surveillance system. Am J Epidemiol. Apr 15, 2014;179(8):1025-1033. [doi: 10.1093/aje/kwu018] [Medline: 24598867]
43. Wheaton AG, Cunningham TJ, Ford ES, Croft JB, Centers for Disease Control and Prevention (CDC). Employment and activity limitations among adults with chronic obstructive pulmonary disease--United States, 2013. MMWR Morb Mortal Wkly Rep. Mar 27, 2015;64(11):289-295. [FREE Full text] [Medline: 25811677]
44. Shiels MS, Chernyavskiy P, Anderson WF, Best AF, Haozous EA, Hartge P, et al. Trends in premature mortality in the USA by sex, race, and ethnicity from 1999 to 2014: an analysis of death certificate data. Lancet. Mar 11, 2017;389(10073):1043-1054. [FREE Full text] [doi: 10.1016/S0140-6736(17)30187-3] [Medline: 28131493]
45. Wheaton AG, Liu Y, Croft JB, VanFrank B, Croxton TL, Punturieri A, et al. Chronic obstructive pulmonary disease and smoking status - United States, 2017. MMWR Morb Mortal Wkly Rep. Jun 21, 2019;68(24):533-538. [FREE Full text] [doi: 10.15585/mmwr.mm6824a1] [Medline: 31220055]

46. Terzikhan N, Verhamme KM, Hofman A, Stricker BH, Brusselle GG, Lahousse L. Prevalence and incidence of COPD in smokers and non-smokers: the Rotterdam study. Eur J Epidemiol. Aug 2016;31(8):785-792. [FREE Full text] [doi: 10.1007/s10654-016-0132-z] [Medline: 26946425]
47. Doo JH, Kim SM, Park YJ, Kim KH, Oh YH, Kim JS, et al. Smoking cessation after diagnosis of COPD is associated with lower all-cause and cause-specific mortality: a nationwide population-based cohort study of South Korean men. BMC Pulm Med. Jul 03, 2023;23(1):237. [FREE Full text] [doi: 10.1186/s12890-023-02533-1] [Medline: 37394482]
48. Perelman J, Alves J, Pfoertner TK, Moor I, Federico B, Kuipers MA, et al. The association between personal income and smoking among adolescents: a study in six European cities. Addiction. Dec 2017;112(12):2248-2256. [FREE Full text] [doi: 10.1111/add.13930] [Medline: 28667824]
49. COPD causes and risk factors. American Lung Association. 2024. URL: https://tinyurl.com/tz5mdhbs [accessed 2024-04-29]
50. Clausen A, Christensen ER, Jakobsen PR, Søndergaard J, Abrahamsen B, Rubin KH. Digital solutions for decision support in general practice - a rapid review focused on systems developed for the universal healthcare setting in Denmark. BMC Prim Care. Dec 14, 2023;24(1):276. [FREE Full text] [doi: 10.1186/s12875-023-02234-y] [Medline: 38097998]
51. Eriksen J, Ebbesen M, Eriksen KT, Hjermitslev C, Knudsen C, Bertelsen P, et al. Equity in digital healthcare - the case of Denmark. Front Public Health. Sep 6, 2023;11:1225222. [FREE Full text] [doi: 10.3389/fpubh.2023.1225222] [Medline: 37744503]
52. Joo JY, Liu MF. Culturally tailored interventions for ethnic minorities: a scoping review. Nurs Open. Sep 09, 2021;8(5):2078-2090. [FREE Full text] [doi: 10.1002/nop2.733] [Medline: 34388862]
53. Radu I, Scheermesser M, Spiess MR, Schulze C, Händler-Schuster D, Pehlke-Milde J. Digital health for migrants, ethnic and cultural minorities and the role of participatory development: a scoping review. Int J Environ Res Public Health. Oct 23, 2023;20(20):6962. [FREE Full text] [doi: 10.3390/ijerph20206962] [Medline: 37887700]
54. Sood A, Petersen H, Liu C, Myers O, Shore XW, Gore BA, et al. Racial and ethnic minorities have a lower prevalence of airflow obstruction than non-Hispanic whites. COPD. 2022;19(1):61-68. [FREE Full text] [doi: 10.1080/15412555.2022.2029384] [Medline: 35099333]
55. Gilkes A, Hull S, Durbaba S, Schofield P, Ashworth M, Mathur R, et al. Ethnic differences in smoking intensity and COPD risk: an observational study in primary care. NPJ Prim Care Respir Med. Sep 04, 2017;27(1):50. [FREE Full text] [doi: 10.1038/s41533-017-0052-8] [Medline: 28871087]
56. Hall WJ, Chapman MV, Lee KM, Merino YM, Thomas TW, Payne BK, et al. Implicit racial/ethnic bias among health care professionals and its influence on health care outcomes: a systematic review. Am J Public Health. Dec 2015;105(12):e60-e76. [doi: 10.2105/AJPH.2015.302903] [Medline: 26469668]
57. Dovidio JF, Penner LA, Albrecht TL, Norton WE, Gaertner SL, Shelton JN. Disparities and distrust: the implications of psychological processes for understanding racial disparities in health and health care. Soc Sci Med. Aug 2008;67(3):478-486. [doi: 10.1016/j.socscimed.2008.03.019] [Medline: 18508171]
58. Williams DR, Sternthal M. Understanding racial-ethnic disparities in health: sociological contributions. J Health Soc Behav. 2010;51 Suppl(Suppl):S15-S27. [FREE Full text] [doi: 10.1177/0022146510383838] [Medline: 20943580]
59. Tucker CM, Marsiske M, Rice KG, Nielson JJ, Herman K. Patient-centered culturally sensitive health care: model testing and refinement. Health Psychol. May 2011;30(3):342-350. [FREE Full text] [doi: 10.1037/a0022967] [Medline: 21553978]
60. Chin JL. Culturally competent health care. Public Health Rep. 2000;115(1):25-33. [FREE Full text] [doi: 10.1093/phr/115.1.25] [Medline: 10968582]
61. Berkowicz R, Palmgren F, Hertel O, Vignati E. Using measurements of air pollution in streets for evaluation of urban air quality — meterological analysis and model calculations. Sci Total Environ. Oct 1996;189-190:259-265. [doi: 10.1016/0048-9697(96)05217-5]
62. Vardoulakis S, Fisher BE, Pericleous K, Gonzalez-Flesca N. Modelling air quality in street canyons: a review. Atmos Environ. Jan 2003;37(2):155-182. [doi: 10.1016/S1352-2310(02)00857-9]
63. Jonidi Jafari A, Charkhloo E, Pasalari H. Urban air pollution control policies and strategies: a systematic review. J Environ Health Sci Eng. Dec 08, 2021;19(2):1911-1940. [FREE Full text] [doi: 10.1007/s40201-021-00744-4] [Medline: 34900316]
64. Ward F, Lowther-Payne HJ, Halliday EC, Dooley K, Joseph N, Livesey R, et al. Engaging communities in addressing air quality: a scoping review. Environ Health. Sep 19, 2022;21(1):89. [FREE Full text] [doi: 10.1186/s12940-022-00896-2] [Medline: 36117163]
65. Rosen LJ, Myers V, Winickoff JP, Kott J. Effectiveness of interventions to reduce tobacco smoke pollution in homes: a systematic review and meta-analysis. Int J Environ Res Public Health. Dec 18, 2015;12(12):16043-16059. [FREE Full text] [doi: 10.3390/ijerph121215038] [Medline: 26694440]
66. Titus AR, Kalousova L, Meza R, Levy DT, Thrasher JF, Elliott MR, et al. Smoke-free policies and smoking cessation in the United States, 2003-2015. Int J Environ Res Public Health. Sep 02, 2019;16(17):3200. [FREE Full text] [doi: 10.3390/ijerph16173200] [Medline: 31480698]
67. Elshawi R, Al-Mallah MH, Sakr S. On the interpretability of machine learning-based model for predicting hypertension. BMC Med Inform Decis Mak. Jul 29, 2019;19(1):146. [FREE Full text] [doi: 10.1186/s12911-019-0874-0] [Medline: 31357998]

68. Hakkoum H, Abnane I, Idri A. Interpretability in the medical field: a systematic mapping and review study. Applied Soft Computing. Mar 2022;117:108391. [FREE Full text] [doi: 10.1016/j.asoc.2021.108391]
69. Meng C, Trinh L, Xu N, Enouen J, Liu Y. Interpretability and fairness evaluation of deep learning models on MIMIC-IV dataset. Sci Rep. May 03, 2022;12(1):7166. [FREE Full text] [doi: 10.1038/s41598-022-11012-2] [Medline: 35504931]
70. Genuer R, Poggi JM, Tuleau-Malot C. Variable selection using random forests. Pattern Recognit Lett. Oct 2010;31(14):2225-2236. [FREE Full text] [doi: 10.1016/j.patrec.2010.03.014]
71. Grömping U. Variable importance assessment in regression: linear regression versus random forest. Am Stat. Nov 2009;63(4):308-319. [doi: 10.1198/tast.2009.08199]
72. Adadi A, Berrada M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access. 2018;6:52138-52160. [doi: 10.1109/access.2018.2870052]
73. Lundberg S, Lee SI. A unified approach to interpreting model predictions. In: Proceedings of the 2017 Annual Conference on Neural Information Processing Systems. 2017. Presented at: NIPS '17; December 4-9, 2017; Long Beach, CA. URL: https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html
74. Kamath SS, Ananthanarayana VS. Semantics-based web service classification using morphological analysis and ensemble learning techniques. Int J Data Sci Anal. Oct 18, 2016;2(1-2):61-74. [doi: 10.1007/s41060-016-0026-x]
75. Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L. Explaining explanations: an overview of interpretability of machine learning. In: Proceedings of the 2018 IEEE 5th International Conference on Data Science and Advanced Analytics. 2018. Presented at: DSAA '18; October 1-3, 2018:80-89; Turin, Italy. URL: https://ieeexplore.ieee.org/document/8631448 [doi: 10.1109/dsaa.2018.00018]
76. Kottemann JE, Remus WE. A study of the relationship between decision model naturalness and performance. MIS Q. Jun 1989;13(2):171-181. [doi: 10.2307/248924]
77. Mannino DM. Chronic obstructive pulmonary disease: definition and epidemiology. Respir Care. Dec 2003;48(12):1185-1191. [FREE Full text] [Medline: 14651759]
78. Black C, Tesfaigzi Y, Bassein JA, Miller LA. Wildfire smoke exposure and human health: significant gaps in research for a growing public health issue. Environ Toxicol Pharmacol. Oct 2017;55:186-195. [FREE Full text] [doi: 10.1016/j.etap.2017.08.022] [Medline: 28892756]
79. Kelley T, Kearney GD. Insights into the environmental health burden of childhood asthma. Environ Health Insights. 2018;12:1178630218757445. [FREE Full text] [doi: 10.1177/1178630218757445] [Medline: 29497308]
80. Wells WD. Lifestyle and psychographics. In: Wells WD, editor. Life Style and Psychographics: Definitions, Uses, and Problems. New York, NY. Marketing Classics Press; 2011.
81. Assael H. A demographic and psychographic profile of heavy internet users and users by type of internet usage. J Adv Res. Oct 12, 2005;45(01):93. [doi: 10.1017/S0021849905050014]
82. Fox X, Purcell K. Chronic disease and the internet. Pew Research Center. URL: https://www.pewresearch.org/internet/2010/03/24/chronic-disease-and-the-internet/ [accessed 2024-04-29]
83. Sujaritpong S, Dear K, Cope M, Walsh S, Kjellstrom T. Quantifying the health impacts of air pollution under a changing climate-a review of approaches and methodology. Int J Biometeorol. Mar 25, 2014;58(2):149-160. [FREE Full text] [doi: 10.1007/s00484-012-0625-8] [Medline: 23354423]
84. Ramanathan V, Feng Y. Air pollution, greenhouse gases and climate change: global and regional perspectives. Atmos Environ. Jan 2009;43(1):37-50. [doi: 10.1016/j.atmosenv.2008.09.063]
85. Barbour E, Deakin EA. Smart growth planning for climate protection. J Am Plann Assoc. Jan 2012;78(1):70-86. [doi: 10.1080/01944363.2011.645272]
86. Boyce JK, Pastor M. Clearing the air: incorporating air quality and environmental justice into climate policy. Clim Change. Aug 6, 2013;120(4):801-814. [doi: 10.1007/S10584-013-0832-2]
87. Cushing L, Blaustein-Rejto D, Wander M, Pastor M, Sadd J, Zhu A, et al. Carbon trading, co-pollutants, and environmental equity: evidence from California's cap-and-trade program (2011-2015). PLoS Med. Jul 10, 2018;15(7):e1002604. [FREE Full text] [doi: 10.1371/journal.pmed.1002604] [Medline: 29990353]
88. Kinney PL. Climate change, air quality, and human health. Am J Prev Med. Nov 2008;35(5):459-467. [doi: 10.1016/j.amepre.2008.08.025] [Medline: 18929972]
89. Strickland E, Andrew NG. Unbiggen AI: the AI pioneer says it's time for smart-sized, "data-centric" solutions to big issues. IEEE Spectrum. 2022. URL: https://tinyurl.com/43bt77tn [accessed 2024-04-29]
90. Liu JC, Pereira G, Uhl SA, Bravo MA, Bell ML. A systematic review of the physical health impacts from non-occupational exposure to wildfire smoke. Environ Res. Jan 2015;136:120-132. [FREE Full text] [doi: 10.1016/j.envres.2014.10.015] [Medline: 25460628]
91. Kamis A, Cao R, He Y, Tian Y, Wu C. Predicting lung cancer in the United States: a multiple model examination of public health factors. Int J Environ Res Public Health. Jun 06, 2021;18(11):6127. [FREE Full text] [doi: 10.3390/ijerph18116127] [Medline: 34204140]
92. Andre M, Sartelet K, Moukhtar S, Andre JM, Redaelli M. Diesel, petrol or electric vehicles: what choices to improve urban air quality in the Ile-de-France region? A simulation platform and case study. Atmos Environ. Nov 2020;241:117752. [doi: 10.1016/j.atmosenv.2020.117752]

93. Lam YF, Fu JS, Wu S, Mickley LJ. Impacts of future climate change and effects of biogenic emissions on surface ozone and particulate matter concentrations in the United States. Atmos Chem Phys. May 23, 2011;11(10):4789-4806. [doi: 10.5194/acp-11-4789-2011]
94. Anenberg SC, Horowitz LW, Tong DQ, West JJ. An estimate of the global burden of anthropogenic ozone and fine particulate matter on premature human mortality using atmospheric modeling. Environ Health Perspect. Sep 2010;118(9):1189-1195. [FREE Full text] [doi: 10.1289/ehp.0901220] [Medline: 20382579]
95. Pope CA, Hansen ML, Long RW, Nielsen KR, Eatough NL, Wilson WE, et al. Ambient particulate air pollution, heart rate variability, and blood markers of inflammation in a panel of elderly subjects. Environ Health Perspect. Mar 2004;112(3):339-345. [FREE Full text] [doi: 10.1289/ehp.6588] [Medline: 14998750]
96. Araujo JA. Particulate air pollution, systemic oxidative stress, inflammation, and atherosclerosis. Air Qual Atmos Health. Nov 10, 2010;4(1):79-93. [FREE Full text] [doi: 10.1007/s11869-010-0101-8] [Medline: 21461032]
97. Dockery DW, Pope CA, Xu X, Spengler JD, Ware JH, Fay ME, et al. An association between air pollution and mortality in six U.S. cities. N Engl J Med. Dec 09, 1993;329(24):1753-1759. [doi: 10.1056/NEJM199312093292401] [Medline: 8179653]
98. Belleudi V, Faustini A, Stafoggia M, Cattani G, Marconi A, Perucci CA, et al. Impact of fine and ultrafine particles on emergency hospital admissions for cardiac and respiratory diseases. Epidemiology. May 2010;21(3):414-423. [doi: 10.1097/EDE.0b013e3181d5c021] [Medline: 20386174]
99. Schraufnagel DE. The health effects of ultrafine particles. Exp Mol Med. Mar 17, 2020;52(3):311-317. [FREE Full text] [doi: 10.1038/s12276-020-0403-3] [Medline: 32203102]
100. Zhang Y, Ding Z, Xiang Q, Wang W, Huang L, Mao F. Short-term effects of ambient PM and PM air pollution on hospital admission for respiratory diseases: case-crossover evidence from Shenzhen, China. Int J Hyg Environ Health. Mar 2020;224:113418. [doi: 10.1016/j.ijheh.2019.11.001] [Medline: 31753527]
101. Ni L, Chuang CC, Zuo L. Fine particulate matter in acute exacerbation of COPD. Front Physiol. Oct 23, 2015;6:294. [FREE Full text] [doi: 10.3389/fphys.2015.00294] [Medline: 26557095]
102. Li T, Hu R, Chen Z, Li Q, Huang S, Zhu Z, et al. Fine particulate matter (PM 2.5): the culprit for chronic lung diseases in China. Chronic Dis Transl Med. Sep 2018;4(3):176-186. [FREE Full text] [doi: 10.1016/j.cdtm.2018.07.002] [Medline: 30276364]
103. Kim CS, Alexis NE, Rappold AG, Kehrl H, Hazucha MJ, Lay JC, et al. Lung function and inflammatory responses in healthy young adults exposed to 0.06 ppm ozone for 6.6 hours. Am J Respir Crit Care Med. May 01, 2011;183(9):1215-1221. [FREE Full text] [doi: 10.1164/rccm.201011-1813OC] [Medline: 21216881]
104. Mudway IS, Kelly FJ. Ozone and the lung: a sensitive issue. Mol Aspects Med. 2000;21(1-2):1-48. [doi: 10.1016/s0098-2997(00)00003-0] [Medline: 10804262]
105. Uysal N, Schapira RM. Effects of ozone on lung function and lung diseases. Curr Opin Pulm Med. Mar 2003;9(2):144-150. [doi: 10.1097/00063198-200303000-00009] [Medline: 12574695]
106. Hendryx M, Luo J, Chojenta C, Byles JE. Air pollution exposures from multiple point sources and risk of incident chronic obstructive pulmonary disease (COPD) and asthma. Environ Res. Dec 2019;179(Pt A):108783. [doi: 10.1016/j.envres.2019.108783] [Medline: 31590000]
107. Chen TM, Kuschner WG, Gokhale J, Shofer S. Outdoor air pollution: nitrogen dioxide, sulfur dioxide, and carbon monoxide health effects. Am J Med Sci. Apr 2007;333(4):249-256. [doi: 10.1097/MAJ.0b013e31803b900f] [Medline: 17435420]
108. Gong Jr H, Linn WS, Clark KW, Anderson KR, Geller MD, Sioutas C. Respiratory responses to exposures with fine particulates and nitrogen dioxide in the elderly with and without COPD. Inhal Toxicol. Mar 06, 2005;17(3):123-132. [doi: 10.1080/08958370590904481] [Medline: 15788373]
109. Lary DJ, Lary T, Sattler B. Using machine learning to estimate global PM2.5 for environmental health studies. Environ Health Insights. 2015;9(Suppl 1):41-52. [FREE Full text] [doi: 10.4137/EHI.S15664] [Medline: 26005352]
110. Tai AP, Mickley LJ, Jacob DJ, Leibensperger EM, Zhang L, Fisher JA, et al. Meteorological modes of variability for fine particulate matter (PM2.5) air quality in the United States: implications for PM2.5 sensitivity to climate change. Atmos Chem Phys. 2012;12(6):3131-3145. [FREE Full text] [doi: 10.5194/acp-12-3131-2012]
111. Peterson GC, Hogrefe C, Corrigan AE, Neas LM, Mathur R, Rappold AG. Impact of reductions in emissions from major source sectors on fine particulate matter–related cardiovascular mortality. Environ Health Perspect. Jan 2020;128(1):017005. [doi: 10.1289/ehp5692]

Abbreviations

AQI: air quality index
CBSA: core-based statistical area
CDC: Centers for Disease Control and Prevention
COPD: chronic obstructive pulmonary disease
GBT: gradient boosted tree
GDP: gross domestic product

LIME: local interpretable model-agnostic explanations
MAE: mean absolute error
ML: machine learning
MLR: multiple linear regression
RMSE: root mean square error
SHAP: Shapley additive explanations
SMAPE: symmetric mean absolute percentage error

Edited by K El Emam, B Malin; submitted 15.03.24; peer-reviewed by GK Gupta, A Wani, S Mao; comments to author 20.04.24; revised version received 09.07.24; accepted 10.07.24; published 29.08.24

Please cite as:
Kamis A, Gadia N, Luo Z, Ng SX, Thumbar M
Obtaining the Most Accurate, Explainable Model for Predicting Chronic Obstructive Pulmonary Disease: Triangulation of Multiple Linear Regression and Machine Learning Methods
JMIR AI 2024;3:e58455
URL: https://ai.jmir.org/2024/1/e58455
doi: 10.2196/58455
PMID: