Chronic Obstructive Pulmonary Disease in the United States: A Comparison of Multiple Linear Regression and Machine Learning Models

Arnold Kamis, Nidhi Gadia, Zilin Luo, Cyndi Ng, Mansi Thumbar

Submitted to: JMIR AI
on: March 15, 2024

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Chronic Obstructive Pulmonary Disease in the United States: A Comparison of Multiple Linear Regression and Machine Learning Models

Arnold Kamis[1] PhD; Nidhi Gadia[1] MS; Zilin Luo[1] MS; Cyndi Ng[1] MS; Mansi Thumbar[1] MS

[1]Brandeis University Waltham US

Corresponding Author:
Arnold Kamis PhD
Brandeis University
415 South St.
Waltham
US

Abstract

Background: Lung disease is a severe problem in the United States. Despite the decreasing rates of cigarette smoking, COPD continues to be health burden in the United States. In this paper, we focus on Chronic Obstructive Pulmonary Disease in the United States from 2016 to 2019.

Objective: We gather a diverse set of data sources to better understand and predict COPD rates at the level of Core-Based Statistical Area in the United States. The objective is to compare linear models with machine learning models to obtain the most accurate and interpretable model of COPD.

Methods: We integrate data from multiple Centers for Disease Control sources and use them to analyze Chronic Obstructive Pulmonary Disease by using different types of methods. We include cigarette smoking, a well-known contributing factor, and race / ethnicity variables because health disparities among different races and ethnicities in the United States are also well-known. The models also include air quality index, education, employment, and economic variables. We fit models with both multiple linear regression and machine learning methods.

Results: The most accurate multiple linear regression model has variance explained = 81.1% and Root Mean Squared Error = 0.73. The most accurate machine learning model has variance explained = 87.1% and Root Mean Squared Error = 0.53. Overall, cigarette smoking and household income are the strongest predictor variables. Hispanic percentage of CBSA, Education, and American Indian / Alaska Native percentage of CBSA are moderately strong predictors.

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 is a Support Vector Machine, which captured non-linearities 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 interventions aimed at decreasing COPD rates. Gaps in understanding the health impacts of air pollution, particularly in relation to climate change, suggest a need for further research to design interventions and improve public health.

(JMIR Preprints 15/03/2024:58455)
DOI: https://doi.org/10.2196/preprints.58455

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Original Paper

Arnold Kamis (first and corresponding author), Brandeis University
akamis@brandeis.edu
ORCID ID: 0000-0002-4222-9730
Associate Professor of Data Analytics

Chronic Obstructive Pulmonary Disease in the United States: A Comparison of Multiple Linear Regression and Machine Learning Models

Chronic Obstructive Pulmonary Disease in the United States: A Comparison of Multiple Linear Regression and Machine Learning Models

Abstract

Background: Lung disease is a severe problem in the United States. Despite the decreasing rates of cigarette smoking, Chronic Obstructive Pulmonary Disease continues to be a health burden in the United States. In this paper, we focus on Chronic Obstructive Pulmonary Disease in the United States from 2016 to 2019.

Objectives: We gather a diverse set of non-personally identifiable information data from public data sources to better understand and predict Chronic Obstructive Pulmonary Disease rates at the level of Core-Based Statistical Area in the United States. Our objective is to compare linear models with machine learning models to obtain the most accurate and interpretable model of Chronic Obstructive Pulmonary Disease.

Methods: We integrate non-personally identifiable information data from multiple Centers for Disease Control sources and use them to analyze Chronic Obstructive Pulmonary Disease by using different types of methods. We include cigarette smoking, a well-known contributing factor, and race / ethnicity variables because health disparities among different races and ethnicities in the United States are also well-known. The models also include air quality index, education, employment, and economic variables. We fit models with both multiple linear regression and machine learning methods.

Results: The most accurate multiple linear regression model has variance explained = 81.1%, Mean Absolute Error = 0.591, and Symmetric Mean Absolute Percentage Error = 9.666. The most accurate machine learning model has variance explained = 85.7%, Mean Absolute Error = 0.456, and Symmetric Mean Absolute Percentage Error = 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 level of Core-Based Statistical Area.

Conclusions: This research highlights the importance of using diverse data sources as well as multiple methods to understand and predict Chronic Obstructive Pulmonary Disease. The most accurate model is a Gradient Boosted Tree, which captured non-linearities 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 Chronic Obstructive Pulmonary Disease 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.

Keywords: Chronic Obstructive Pulmonary Disease; cigarette smoking; ethnic and racial differences; machine learning; multiple linear regression; household income

Introduction

Lung disease is a severe problem in the United States. According to the Centers for Disease Control and Prevention, asthma is responsible for at least 3,000 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, e.g., cigarette smoking, and others not under individual control, e.g., 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 vulnerable populations. 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 of days when it is dangerously poor air quality outside, advising them to stay indoors or to avoid strenuous outdoor exercise. The effectiveness of such apps is mixed 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. We then 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 a significant literature on factors contributing to COPD, including a wide variety of environmental, economic, and demographic variables; the etiology of COPD is multifactorial, with smoking by far 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]. Additionally, occupational exposures, e.g., to coal dust, arsenic, diesel fumes, or to home exposures, e.g., 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 co-pollutants are associated with decreased lung function and can lead to COPD. The loss can range from mild, e.g., allergies, to severe, i.e., mortality. Air quality varies widely throughout the United States because of pollutants and co-pollutants, and climate change may be worsening it, particularly for vulnerable populations [9]. Aeroallergens have been found to increase the incidence of COPD [5]. Health disparities due to poor quality air and other stressors are well-known [10-12]. Ambient air pollution in poorer neighborhoods will tend to be exacerbated by additional co-pollutants, heat stress, and aeroallergens. Air Quality Index (AQI) includes the totality of pollutants and co-pollutants.

Machine learning 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, a general protocol was developed 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 [14]. 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 have been used to predict lung function values from computed tomography (CT) images obtained from COPD and non-COPD patients [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 to public health use cases in general and COPD in particular. The one that ultimately works best in a given situation depends on many factors.

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

We have three general expectations in our models 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 / ethnic demographics.

Methods

This paper uses multiple linear regression and machine learning methods to predict COPD at the level of Core-Based Statistical Area (CBSA) [25]. There are, at the time of this study, 388 metropolitan statistical areas and 541 micropolitan statistical areas in the United States. The data sources were obtained from three official United States agency data repositories, specifically from the Centers for Disease Control. 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 (NULL/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	URL
National Center for Health Statistics	https://www.cdc.gov/nchs
Chronic Disease Indicators Data	https://chronicdata.cdc.gov
US Chronic Disease Indicator, Stratification values	https://www.cdc.gov/cdi

Data was collected for all CBSA that were available in 2016-2019. All data obtained from the CDC was contributed voluntarily at the individual level and aggregated to remove all personally identifiable information [26].

COPD rates are for 2019 whereas all the predictor variables are averaged over the timespan 2016-2018. As such, the models obtained are predictive over time. The data collection result was 517 out of the 929 core-based statistical areas, 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 model all factors as random variables directly contributing to COPD, which is measured as the proportion (percentage) of the population having COPD. Race and ethnicity are 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.

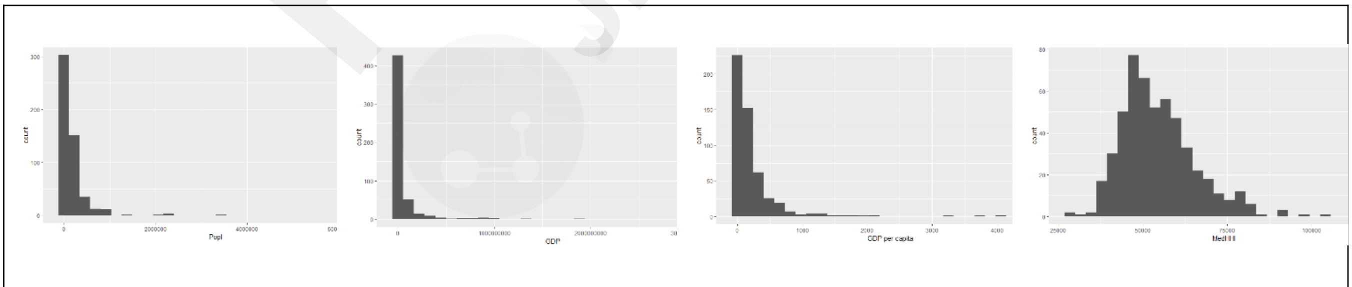
Table 2: Main Variables and Descriptive Statistics, average within CBSA

	Year(s)	Min.	1stQuartile	Median	Mean	3rdQuartile	Max.
population	2016-18	7,351	48,763	96,811	191,892	180,484	6,633,096
GDP	2016-18	$447,355	$2,562,704	$13,126,907	$64,223,036	$39,046,120	$3,218,209,695
Median HH Income	2016-18	$27,842	$46,867	$52,632	$54,736	$60,494	$119,332
GDP per capita	2016-18	$16.86	$47.83	$100.07	$253.77	$277.17	$4,731.50
air quality index	2016-18	9.00	34.00	38.67	38.02	43.00	95.00
smoking rate	2016-18	8.41	15.33	17.12	17.29	19.28	29.59
poverty rate (all ages)	2016-18	3.87	10.92	13.80	14.36	17.12	35.56

	Year(s)	Min.	1stQuartile	Median	Mean	3rdQuartile	Max.
unemployment rate	2016-18	1.97	3.67	4.52	4.71	5.43	20.93
education rate	2016-18	8.77	17.94	22.91	24.22	27.96	65.75
White	2016-18	22.1%	78.6%	87.6%	84.6%	92.8%	100.0%
Black	2016-18	0.3%	1.5%	4.0%	9.3%	12.5%	100.0%
AI_or_AN	2016-18	0.1%	0.4%	0.7%	2.0%	1.7%	45.9%
Asian	2016-18	0.2%	0.9%	1.6%	2.8%	3.0%	42.8%
NH_or_PI	2016-18	0.0%	0.1%	0.1%	0.3%	0.2%	12.9%
Hispanic	2016-18	0.9%	3.9%	7.0%	13.3%	14.9%	95.5%
COPD rate	2019	3.2	5.7	6.7	6.871	7.9	15

Notes: AI_or_AN: American Indian or Alaska Native; NH_or_PI: Native Hawaiian or Other Pacific Islander

We observe in Figure 1 that some variables (Population, GDP, GDP per capita, and median Household Income) are skewed in their distribution.



Population GDP GDP Per Capita Median Household Income

Figure 1: Population, GDP, GDP Per Capita, and Median Household Income.

We therefore make a log transformation of them (logPopl, logGDP, logGDPpc, logHHI) to make them less skewed, and show a heatmap of correlations of them with the other variables in Figure 2.

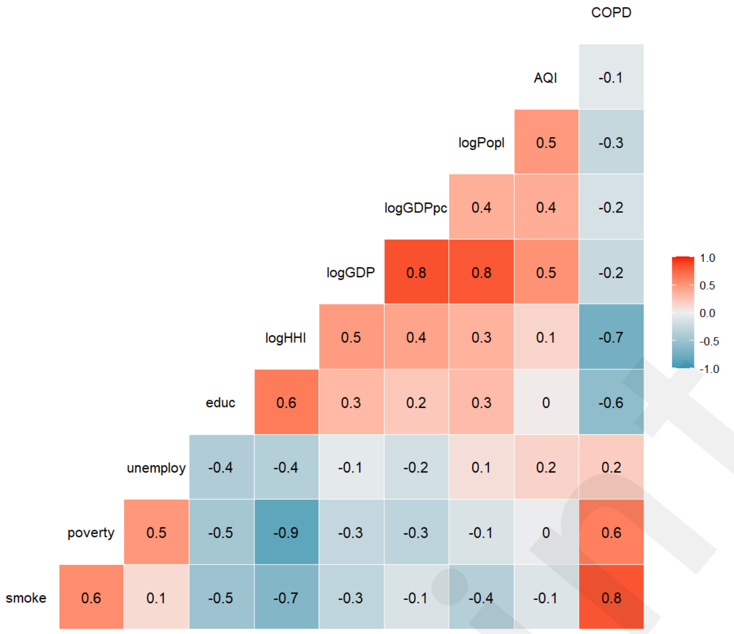


Figure 2: Correlations among main variables.

We see a range of correlations, from very negative (dark blue) to very positive (dark red). We see in the right-most column 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 household income, logged) to slightly negative (log of GDP, log of Population, log of GDP per capita, and Air Quality Index). Given these correlations, we are likely to find good predictive models, but we need to check for multi-collinearity 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. We chose in this paper to include cigarette smoking, accounting for it in our models, but also examine other factors in order to compare the magnitudes of influence among the various factors. We aim to model a variety of factors, including cigarette smoking, to arrive at the model that predicts COPD with the greatest accuracy.

Statistical Analysis

Our multiple linear regression baseline model in R 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	Std. Error	t value	*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

	Estimate	Std. Error	t value	*P* value
AI_or_AN_percentage	-2.9000	0.778	-3.726	<.001
Black_percentage	-1.2700	0.356	-3.558	<.001
NH_or_PI_percentage	-15.8000	4.430	-3.558	<.001
Log_GDP	0.0741	0.035	2.126	0.034
White_percentage	-0.7380	0.358	-2.060	0.04
Unemployment_Rate	0.0456	0.023	1.993	0.047
Asian_percentage	2.4800	1.250	1.988	0.047
Log_Population	0.0899	0.055	1.626	0.105
Air_Quality_Index	0.0003	0.003	0.098	0.922
Residual standard error: 0.658 on 503 degrees of freedom				
Multiple R-squared: 0.8152, Adjusted R-squared: 0.8105				

The model has F-statistic 170.7 on 13 predictor variables and 503 DF, *P*<.001. The Variance Inflation Factors were checked, with all values less than 5, indicating low multi-collinearity.

There are seven predictors of high statistical significance: Smoking Rate, Black percentage, NH_or_PI percentage, AI_or_AN 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 percentage increase in the COPD rate. The other six highly significant predictors have a negative association. Every percentage increase in Log of household income lowers the COPD rate by 2.81 percent. Nearly as strong is the Hispanic percentage; every percentage increase corresponds to a drop of 2.39 percentage in COPD rate. AI_or_AN is a bit stronger in its coefficient estimate; every percentage point increase corresponds to a drop of 2.9 percent in COPD rate. Much stronger, every percentage point increase in NH_or_PI corresponds to a drop of 15.8 percent. Every percentage point increase in Black percentage corresponds to a drop of 1.27 percent. Education rate has a strongly statistically significant relationship, but a small percentage point impact: every percentage increase corresponds to a decrease of 0.0334 percentage in COPD rate. The remaining four 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. We argue that one should not stop there, however, because an ML model can capture substantial variance from non-linear 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, i.e., decision trees. ML models can add nuance to those decision trees based on the captured non-linearities. Although an adjusted R-Squared of 0.8105 looks quite strong, we can perhaps do better with ML methods [18-21].

The seven ML methods evaluated in this paper are Lasso Regression, Ridge Regression, Generalized Additive Model, Support Vector Machine, Artificial Neural Network, Random Forest, and Gradient Boosted Tree. These methods were selected for their known strengths in minimizing Errors of Bias and/or Errors of Variance, i.e., 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 4.2.3. We summarize each method in terms of its main pros and cons:

- Lasso Regression (L1 regularization): a multiple linear regression 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 zero, effectively performing feature selection by excluding less important variables from the model. This reduces model complexity and minimizes multi-collinearity. This is a standard refinement of multiple linear regression. (R package glmnet)
- Ridge Regression (L2 regularization): a multiple linear regression technique that adds a penalty term to the objective function to reduce the coefficients of less important predictors and guard against over-weighting 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 multi-collinearity. This is a standard refinement of multiple linear regression. (R package ridge)
- Generalized Additive Model: a nonparametric generalization of multiple linear regression, 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 multiple linear regression, all the terms are added together. Although overfitting can occur, regularization and cross-validation help to minimize it. (R package mgcv).
- Support Vector Machine: 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 for 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: generalization of multiple linear regression with hidden layer(s) of nodes between input and output nodes; 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 one or more 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: 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 Tree: 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 machine learning approach follows best practices. We randomly partition the dataset into train (60%), cross-validate (20%), and test (20%) subsets. We check for outliers, multi-collinearity, and target leakage to ensure valid models [27].

Machine Learning Model Results

In Table 4, we describe the results of the ML models of COPD by various accuracy metrics. For the accuracy metrics, we used three standard measures of predictive accuracy in addition to variance explained (adjusted R-squared): Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Symmetric Mean Absolute Percentage Error (SMAPE) [28, 29]. We performed a grid search over all the main numeric parameters for a given method to find the optimal combination of parameter values [30]. 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 **boldface**.

Table 4: Machine Learning Models vs. Multiple Linear Regression

		RMSE			MAE			SMAPE		
method	adj. R^2	Train	CV	Test	Train	CV	Test	Train	CV	Test
Gradient Boosted Tree (xgboost, loss function = least squares, learning rate = 0.05, max tree depth = 10)	0.857	0.550	0.598	0.557	0.433	0.445	**0.456**	6.473	6.543	**6.956**
Support Vector Machine (Nystroem kernel, 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 (max trees = 500, max depth = none, max 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, 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, max bins = 100, 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

The ML methods are superior to Multiple Linear Regression on most metrics. Support Vector Machine is the best on adjusted R^2 and RMSE, slightly superior to Gradient Boosted Tree, but Gradient Boosted Tree is superior by a larger margin on MAE and SMAPE. Therefore, we choose Gradient Boosted Tree (GBT) as the best overall method. In Figure 3, we show the variable importance plot for the Gradient Boosted Tree 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 [31].

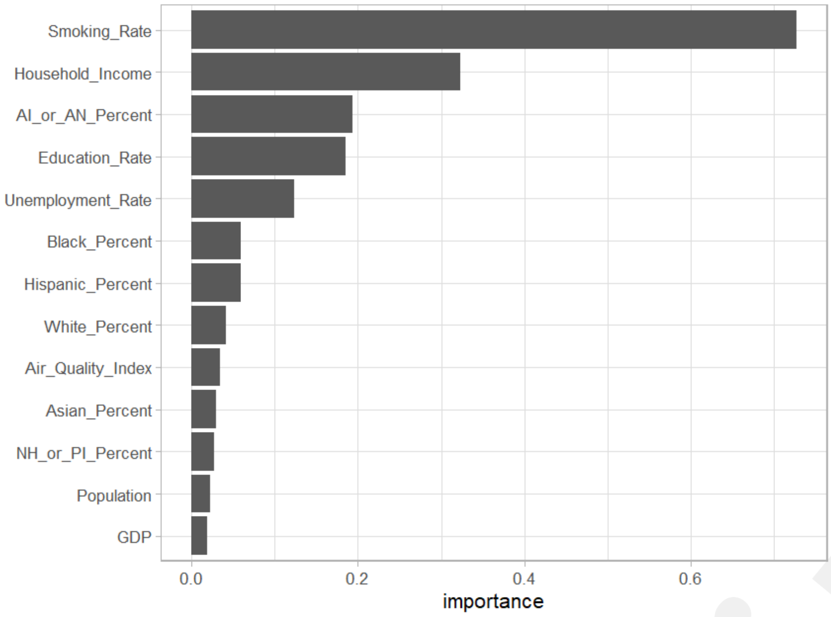


Figure 3: Importance of model variables for Gradient Boosted Tree.

The top five variables in terms of impact are 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 is sixth, Hispanic percentage is seventh, and there is only a small impact from the remaining variables: White percentage, Air Quality Index, 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 rank importance. Hispanic percentage dropped from 3rd to 7th rank, American Indian or Alaska Native percentage rose from 5th to 3rd 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 4 shows the lift plot, and Figure 5 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 is calculated and plotted. The predictive residual plot shows the differences between observed and predicted values.

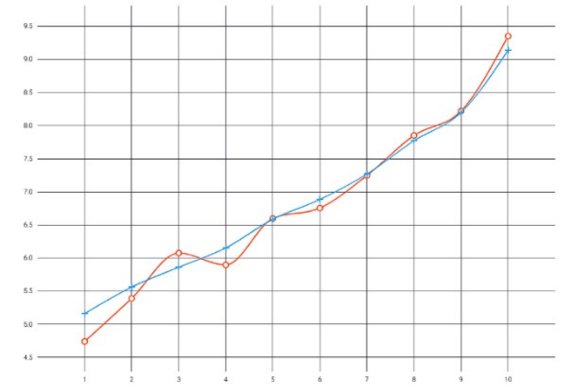


Figure 4. Lift plot showing COPD rate as a function of 10 decile bins. Predicted Values (blue), Actual Values (red)

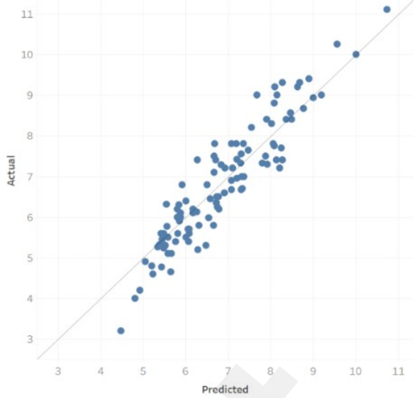


Figure 5. Prediction residuals.

In addition to the variable importance plot, other plots have been used to gain understanding of ML models: LIME models and SHAP plots [32-34]. We chose SHAP plots because they are based on a game-theoretic foundation, showing every combination of the variables in the model and how they work together to predict the outcome variable. Figure 6 shows the SHAP plot for all the Gradient Boosted Tree's variables.

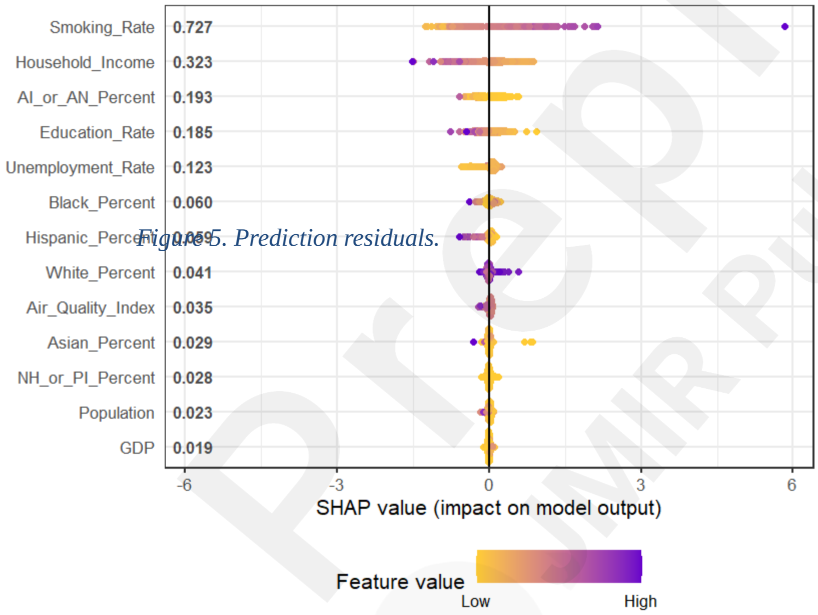


Figure 6. SHAP values for all features (variables).

The top five variables (Smoking Rate, Household_Income, AI_or_AN_percentage, Education Rate, and Unemployment Rate) have substantially more impact on COPD percentage than the remaining variables. We show the top five as well as the next four as individual SHAP plots of the GBT in Figure 7. All nine of the plots show significant non-linearities.

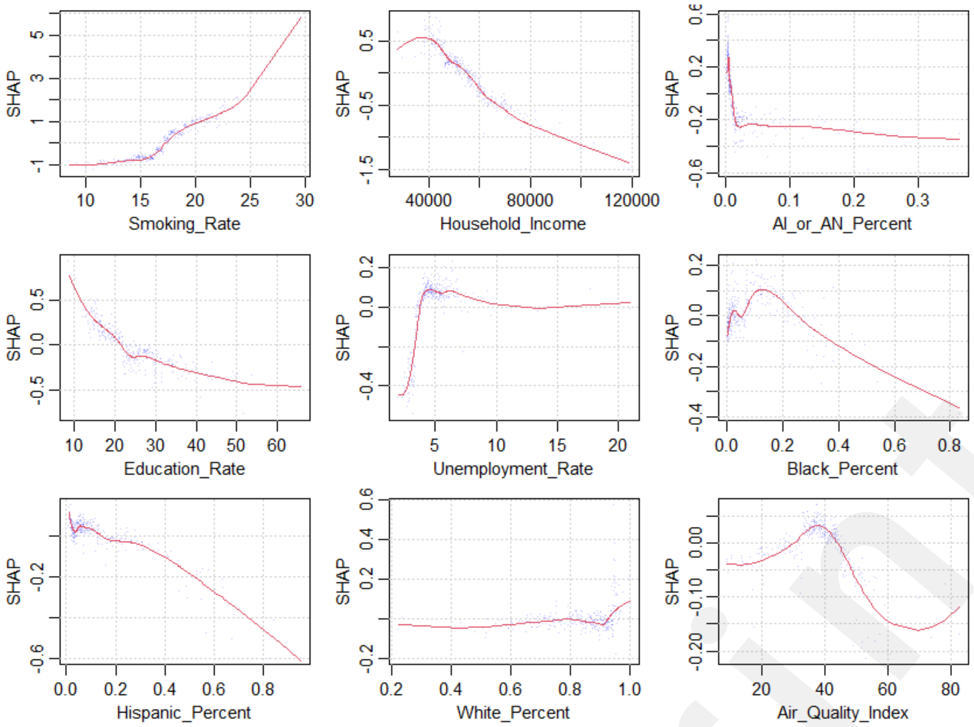


Figure 7. SHAP plots for the nine most important variables.

Smoking has the biggest impact: the higher the smoking rate, the higher the COPD rate, in a steeply curved — almost exponential — relationship. Median household income has 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, healthcare access, higher quality healthcare (prevention and/or treatment), or lower occupational / home exposure, e.g., gas stoves. The next variable is American Indian or Alaska Native percentage, indicating a negative, but non-linear relationship to 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.

The next variable, Education Rate has 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, healthcare access, and/or quality of healthcare, or lower occupational/home exposure [35]. The next variable is Unemployment rate, with a sharply positive but then flat relationship to COPD rate. The next variable is Black percentage, with an initial positive relationship to COPD rate, but then a reversal to a negative, linear relationship to COPD rate.

The next variable, Hispanic percentage, shows a negative linear relationship to COPD rate. This represenats a significant protective influence shown for people in the Hispanic community, which is consistent with the literature [36-42]. The next variable is White percentage, showing a slightly negative relationship to COPD rate. Finally, the last variable is Air Quality Index (higher is 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.

Principal Results

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 machine learning model, but it was smaller than expected.
3. There were substantial racial/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 to 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 [43]. 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 [44].

Three of the four next most important variables, in terms of impact in our GBT, are socio-economic: Household Income (#2), Education Rate (#4) and Unemployment Rate (#5). In the MLR, we found that Household Income (logged) has the second highest impact. In the GBT, Household Income had the second highest impact, but the tipping point was around $40,000, after which higher income has a linear, negative relationship with COPD. Education Rate has a strongly negative, curvilinear relationship with COPD. Unemployment Rate has a sharply positive relationship with COPD, but then peaks at 5% unemployment, after which it plateaus.

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 six European cities found that disposable income was positively associated with smoking [45]. Conversely, lower socioeconomic status is associated with higher COPD prevalence, because in addition to lower education and income, there may be environmental pollutants, occupational hazards, and/or barriers to COPD screening / diagnosis / treatment [46]. In contrast with the literature, our SHAP plots show mostly non-linear relationships with COPD. Household Income shows a tipping point at $40,000, after which the negative relationship with COPD is nearly linear.

Ethnic / Racial variables account for three of the top seven variables in the GBT: American Indian or Alaska Native percentage (#3), Black percentage (#6), and Hispanic percentage (#7). The greater the size of those minority populations, the lower the COPD rate. Our SHAP plots show significant tipping points (non-linearities) for American Indian or Alaska Native percentage and Black percentage, and a mostly linear relationship for Hispanic percentage. Consistent with the literature, all three 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 non-linear impact on larger Black populations. Higher Education Rate and Lower Unemployment Rate are also protective, whereas Air Quality Index shows mixed effects. These results have implications for private healthcare practitioners, public healthcare officials, and healthcare policymakers who aim to reduce COPD rates. Such policies and programs should not assume high digital literacy [47, 48]. System designers could use 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 / ethnic groups, members of the various communities should be included in the design process [49, 50]. In sum, the user interface should exhibit high ease-of-use — employing 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 minorities, particularly African Americans and Hispanics, had a lower prevalence of airflow obstruction than non-Hispanic whites, even after adjusting for smoking status and other risk factors [51]. This finding was supported by another study that observed lower COPD risk in ethnic minority groups compared to whites, despite similar smoking intensities [52]. A larger minority population means a larger peer support network for prevention / cessation of smoking and a larger peer community to recommend COPD screening / diagnosis / treatment, which is particularly useful in a healthcare system that has implicit racial/ethnic bias [47, 53].

There are varying levels of patient trust and implicit bias in healthcare practitioners themselves [54], 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 [55]. Healthcare 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. Healthcare practitioners and educators should communicate to different populations in culturally sensitive ways [56, 57].

Educational materials and behavior change strategies may need to be customized according to different risk factors, beliefs, preferences, and techno-graphics of different sub-populations [47, 48]. On a basic level, people with lower levels of education or household income could be directed via phone geo-location to their local healthcare and to their community leaders for in-person guidance and support. Those local leaders could then advise them of local smoking cessation programs and apps / 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.

Air Quality Index was not significant in the multiple linear regression, but it was significant in the Gradient Boosted Tree, albeit not as strongly as we expected. It could be that the AQI is more of a diffuse, macro-level environmental factor that fluctuates over time, making some CBSAs worse on average, but with wide volatility, e.g., as weather and wind directions change [58, 59]. 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 vulnerable communities, i.e., communities of lower education, higher unemployment, and lower household income. Subsidies for households in vulnerable communities to convert to more efficient, cleaner home heating/cooling methods would improve their home's air quality at a lower cost [60]. Research suggests that engaging communities in targeting their air quality issues can lead to more positive outcomes in both air quality and public health [61-63].

Discussion

There is a small but growing body of research that uses ML models in healthcare / 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 a recent review:

- Elshawi et. al (2019) compared model-agnostic explanations using two techniques, Local Interpretable Model-Agnostic Explanations (LIME) and Shapley values, to interpret a machine learning model for predicting hypertension risk. LIME uses small subsets of the data, which may be idiosyncratic, to provide intuitive explanations, i.e., rules. Shapley values are more theoretically sound and global, using all the available data and therefore less idiosyncratic than LIME, but they do not provide LIME's simple, linear explanations [64].
- Hakkoum et. al (2022) conducted an extensive literature review of machine Learning interpretability in medicine published between 1994 and 2020. The review found that there is no consensus on evaluation metrics or frameworks to assess the quality and utility of the interpretability methods [65]. The highest performing ML models do not translate easily into clinical rules.
- Meng et. al (2022) reviewed the interpretability and fairness evaluation of deep learning models on MIMIC-IV dataset, a large, publicly available benchmark for developing and evaluating the interpretability of high-performing machine learning models, which use sensitive demographic features. The review found that existing interpretation methods, e.g., 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 healthcare. There are always tradeoffs between accuracy and interpretability / explainability. We chose to use Shapley values, because they represent the frontier in explainability, and they are similar to interpreting a multiple regression, interpreting one variable at a time, without making the assumptions of linear models. Additionally, Shapley values allow for non-linear 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 to COPD.

Our best multiple linear regression model has variance explained = 81.1%, MAE = 0.591, and SMAPE = 9.666. Our best machine learning model is a Gradient Boosted Tree, with variance explained of 85.7%, MAE = 0.456 and SMAPE = 6.956. The GBT explains the vast majority of the variance — 4.6% more than the best multiple linear regression — with far less predictive error. We can see that the Gradient Boosted Tree's SMAPE (6.956) is 28% lower than that of the MLR's SMAPE (9.666). Similarly, the GBT's RMSE is 26% lower than the MLR's RMSE and its MAE is 23% lower than that of the MLR. Real-world predictive accuracy should be similar to that found in the Test dataset because the Test data was never used in the GBT's model development.

Our GBT performs strongly on the test data, with very little performance deterioration on the test data vs. 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 [31, 66, 67] and SHAP plots [68, 69]. The 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" [70-72], thus preserving interpretability and actionability, in addition to adding non-linear nuance.

Limitations and Future Directions

This research has a few limitations. The data was obtained from 517 out of the 929 core-based statistical areas. We assumed that this was an adequate sample, i.e., 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, i.e., 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 [73-75]. Future research could develop separate models stratified by demographic variables such as race / ethnicity, assuming there is sufficient data for each categorical class. There could also be geopolitical variations in terms of population density as well as demographics, psychographics [76], and technographics [77, 78].

Future data collection could focus on understanding racial / 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 healthcare system and population? Is it related to successfully tailored smoking prevention / 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 of our models were structured as direct effects. We applied multiple linear regression and machine learning methods with data from CBSAs, which have significant variation in terms of healthcare access and quality. Using these models as a foundation, we should recognize the interconnectedness (direct, indirect, and interactive) of pollutants and co-pollutants 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, e.g., showing how asthma may lead to COPD or to asthma-COPD Overlap [73].

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 [79]. Future research directions could focus on modeling the direct and indirect links between environmental exposures and COPD. Based on those results, we could design interventions, e.g., 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 [80-84], contributing indirectly to COPD. Future research could collect data on new, additional variables pertaining to climate change [85]. Wildfires, which are increasingly common, produce more carcinogens in the air, including high levels of particulate matter. This can directly decrease air quality or co-pollute with other ambient pollutants [86]. These problems have been shown to increase the odds of lung cancer [87], 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, 87-90]. Coarse, fine and ultrafine particulate matter have been studied extensively and linked to systemic oxidative stress, inflammation [91], atherosclerosis [92], and mortality [93] in the United States [94, 95] as well as in China [96-98]. Tropospheric ozone (O_3) exposure by itself has been linked to impaired lung function and increased COPD-related hospital admissions [99-101].

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 [102, 103] and in the elderly [104]. In sum, data pertaining to ambient pollution, e.g., Particulate Matter, Sulfur Dioxide, and Carbon Monoxide, could be useful additional co-pollutant data to include in future models [6, 82-84, 87, 105-107].

Conclusions

Our novel contributions in this paper include 1) integrating multiple publicly available Centers for Disease Control data sources, 2) development of highly accurate models using linear and non-linear methods, and 3) interpretation of the variable impacts for the best model. Smoking is the number one 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). Race / 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 machine learning methods in general and a Gradient Boosted Tree in particular, which produced a highly accurate model of COPD rates. The computational complexity of a Gradient Boosted Tree enables it to obtain high accuracy, but healthcare policymakers may be reluctant to adopt it unless they can obtain a rule-based explanation. Also, 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 algorithm is audited, and it cannot be explained, there may be serious legal or financial consequences [68]. That is why it is important to have explainable models, to open the “black-box,” rendering them interpretable and actionable [71, 72]. This research shows that it is possible to do so.

Acknowledgements

We 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. We also thank anonymous reviewers at the International Business School, Brandeis University, for their helpful comments.

Conflicts of interest

The authors have no conflicts of interest to declare.

Abbreviations

AI_or_AN: American Indian or Alaska Native

AQI: Air Quality Index

CBSA: Core-Based Statistical Area

COPD: Chronic Obstructive Pulmonary Disease

GDP: Gross Domestic Product

HHI: Household Income

MAE: Mean Absolute Error

NH_or_PI: Native Hawaiian or Other Pacific Islander

RMSE: Root Mean Square Error

SMAPE: Symmetric Mean Absolute Percentage Error

SVM: Support Vector Machine

Data Availability Statement: n/a

Author Contributions: The authors contributed equally to this project.

Funding: Not applicable.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

References

1. Diaz-Guzman, E, M Khosravi and DM Mannino, Asthma, chronic obstructive pulmonary disease, and mortality in the US population. COPD: Journal of Chronic Obstructive Pulmonary Disease; 2011. **8**(6): p. 400-407. doi: 10.3109/15412555.2011.611200.
2. Hardin, M, EK Silverman, RG Barr, NN Hansel, JD Schroeder, BJ Make, JD Crapo, CP Hersh and C Investigators, The clinical features of the overlap between COPD and asthma. Respiratory Research; 2011. **12**. doi: 10.1186/1465-9921-12-127.
3. Peden, DB, Air pollution in asthma: effect of pollutants on airway inflammation. Annals of Allergy, Asthma & Immunology; 2001. **87**(6): p. 12-17. doi: 10.1016/S1081-1206(10)62334-4.
4. Wong-Parodi, G, MB Dias and M Taylor, 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; 2018. **6**(3): p. e48. doi: 10.2196/mhealth.8273.
5. Iribarren, SJ, TO Akande, KJ Kamp, D Barry, YG Kader and E Suelzer, Effectiveness of Mobile Apps to Promote Health and Manage Disease: Systematic Review and Meta-analysis of Randomized Controlled Trials. JMIR Mhealth Uhealth; 2021. **9**(1): p. e21563. doi: 10.2196/21563.
6. Valavanidis, A, T Vlachogianni and K Fiotakis, Tobacco Smoke: Involvement of Reactive Oxygen Species and Stable Free Radicals in Mechanisms of Oxidative Damage, Carcinogenesis and Synergistic Effects with Other Respirable Particles. International Journal

of Environmental Research and Public Health; 2009. **6**(2): p. 445-462. doi: 10.3390/ijerph6020445.

7. Berend, N, Contribution of air pollution to COPD and small airway dysfunction. Respirology; 2016. **21**(2): p. 237-244. doi: 10.1111/resp.12644.
8. Lissåker, CTK, EO Talbott, H Kan and X Xu, Status and determinants of individual actions to reduce health impacts of air pollution in US adults. Archives of Environmental & Occupational Health; 2016. **71**(1): p. 43-48. doi: 10.1080/19338244.2014.988673.
9. Dransfield, MT and WC Bailey, COPD: Racial Disparities in Susceptibility, Treatment, and Outcomes. Clinics in Chest Medicine; 2006. **27**(3): p. 463-471. doi: 10.1016/j.ccm.2006.04.005.
10. Duru, OK, NT Harawa, D Kermah and KC Norris, Allostatic Load Burden and Racial Disparities in Mortality. Journal of the National Medical Association; 2012. **104**(1): p. 89-95. doi: 10.1016/S0027-9684(15)30120-6.
11. Alexander, GR, MS Wingate, D Bader and MD Kogan, The increasing racial disparity in infant mortality rates: Composition and contributors to recent US trends. American Journal of Obstetrics and Gynecology; 2008. **198**(1): p. 51.e1-51.e9. doi: 10.1016/j.ajog.2007.06.006.
12. Woolf, SH, RE Johnson, GE Fryer, G Rust and D Satcher, The Health Impact of Resolving Racial Disparities: An Analysis of US Mortality Data. American Journal of Public Health; 2004. **94**(12): p. 2078-2081. doi: 10.2105/AJPH.94.12.2078.
13. Budd, J, BS Miller, EM Manning, V Lampos, M Zhuang, M Edelstein, G Rees, VC Emery, MM Stevens, N Keegan, MJ Short, D Pillay, E Manley, IJ Cox, D Heymann, AM Johnson and RA McKendry, Digital technologies in the public-health response to COVID-19. Nat Med; 2020. **26**(8): p. 1183-1192. doi: 10.1038/s41591-020-1011-4.
14. Zicari, RV, J Brusseau, SN Blomberg, HC Christensen, M Coffee, MB Ganapini, S Gerke, TK Gilbert, E Hickman, E Hildt, S Holm, U Kühne, VI Madai, W Osika, A Spezzatti, E Schnebel, JJ Tithi, D Vetter, M Westerlund, R Wurth, J Amann, V Antun, V Beretta, F Bruneault, E Campano, B Düdder, A Gallucci, E Goffi, CB Haase, T Hagendorff, P Kringen, F Möslein, D Ottenheimer, M Ozols, L Palazzani, M Petrin, K Tafur, J Tørresen, H Volland, and G Kararigas, On Assessing Trustworthy AI in Healthcare. Machine Learning as a Supportive Tool to Recognize Cardiac Arrest in Emergency Calls. Frontiers in Human Dynamics; 2021. **3**. doi:
15. Wenstrup, J, JD Havtorn, L Borgholt, SN Blomberg, L Maaloe, MR Sayre, H Christensen, C Kruuse and HC Christensen, A retrospective study on machine learning-assisted stroke recognition for medical helpline calls. NPJ digital medicine; 2023. **6**(1): p. 235. doi:
16. Wu, C-T, G-H Li, C-T Huang, Y-C Cheng, C-H Chen, J-Y Chien, P-H Kuo, L-C Kuo and F Lai, 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 and uHealth; 2021. **9**(5): p. e22591. doi:
17. Peng, J, C Chen, M Zhou, X Xie, Y Zhou and C-H Luo, A machine-learning approach to forecast aggravation risk in patients with acute exacerbation of chronic obstructive pulmonary disease with clinical indicators. Scientific reports; 2020. **10**(1): p. 3118. doi:
18. Moll, M, D Qiao, EA Regan, GM Hunninghake, BJ Make, R Tal-Singer, MJ McGeachie, PJ Castaldi, RSJ Estepar and GR Washko, Machine learning and prediction of all-cause mortality in COPD. Chest; 2020. **158**(3): p. 952-964. doi:
19. Lee, Y, E Kim, KJ Chae, CH Lee, GY Jin, SR Kim and J Choi, Machine Learning Predicts Computed Tomography (CT)-based Normal Regional Lung Function Distribution in Asthma and Chronic Obstructive Pulmonary Disease (COPD), in *B68. A DIFFERENT POINT OF VIEW: LUNG IMAGING IN COPD*. 2023, American Thoracic Society. p. A4004-A4004.
20. Estépar, RSJ, Artificial Intelligence in COPD: New Venues to Study a Complex Disease. Barc Respir Netw Rev; 2020. **6**(2): p. 144-160. doi: 10.23866/BRNRev:2019-0014.

21. Fortis, S, Y Gao, AK Baldomero, MV Sarrazin and PJ Kaboli, Association of rural living with COPD-related hospitalizations and deaths in US veterans. Scientific Reports; 2023. **13**(1): p. 7887. doi: 10.1038/s41598-023-34865-7.
22. Freimuth, VS, SC Quinn, SB Thomas, G Cole, E Zook and T Duncan, African Americans' views on research and the Tuskegee Syphilis study. Social Science & Medicine; 2001. **52**(5): p. 797-808. doi: 10.1016/S0277-9536(00)00178-7.
23. Vozoris, NT and MB Stanbrook, Smoking prevalence, behaviours, and cessation among individuals with COPD or asthma. Respiratory Medicine; 2011. **105**(477-484). doi: 10.1016/j.rmed.2010.08.011.
24. Ezzati, M, AB Friedman, SC Kulkarni and CJL Murray, The Reversal of Fortunes: Trends in County Mortality and Cross-County Mortality Disparities in the United States. PLOS Medicine; 2008. **5**(4): p. e66. doi: 10.1371/journal.pmed.0050066.
25. Bureau, USC. Housing Patterns and Core-Based Statistical Areas. 2021; Available from: https://www.census.gov/topics/housing/housing-patterns/about/core-based-statistical-areas.html.
26. CDC. Office of Public Health Data, Surveillance, and Technology (OPHDST). 2024 January 26, 2024]; Available from: https://www.cdc.gov/about/divisions-offices/ophdst.html.
27. Kaufman, S, S Rosset, C Perlich and O Stitelman, Leakage in data mining: Formulation, detection, and avoidance. ACM Transactions on Knowledge Discovery from Data; 2011(15): p. 556–563. doi: 10.1145/2382577.2382579.
28. Hyndman, RJ and AB Koehler, Another look at measures of forecast accuracy. International Journal of Forecasting; 2006. **22**(4): p. 679-688. doi: 10.1016/j.ijforecast.2006.03.001.
29. Willmott, CJ and K Matsuura, Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Research; 2005. **30**: p. 79–82. doi: 10.3354/cr030079.
30. Fayed, HA and AF Atiya, Speed up grid-search for parameter selection of support vector machines. Applied Soft Computing; 2019. **80**: p. 202-210. doi: 10.1016/j.asoc.2019.03.037.
31. Greenwell, BM and BC Boehmke, Variable Importance Plots-An Introduction to the vip Package. R J.; 2020. **12**(1): p. 343-366. doi: 10.32614/RJ-2020-013.
32. Antwarg, L, RM Miller, B Shapira and L Rokach, Explaining anomalies detected by autoencoders using Shapley Additive Explanations. Expert Systems with Applications; 2021. **186**: p. 115736. doi: https://doi.org/10.1016/j.eswa.2021.115736.
33. Garreau, D and U Luxburg. Explaining the explainer: A first theoretical analysis of LIME. 2020. PMLR.
34. Gramegna, A and P Giudici, SHAP and LIME: an evaluation of discriminative power in credit risk. Frontiers in Artificial Intelligence; 2021. **4**: p. 752558. doi:
35. Hetlevik, Ø, H Melbye and S Gjesdal, GP utilisation by education level among adults with COPD or asthma: a cross-sectional register-based study. NPJ Prim Care Respir Med; 2016. **26**: p. 16027. doi: 10.1038/npjpcrm.2016.27.
36. daCosta DiBonaventura, M, R Paulose-Ram, J Su, M McDonald, KH Zou, J-S Wagner and H Shah, The Impact of COPD on Quality of Life, Productivity Loss, and Resource Use among the Elderly United States Workforce. COPD: Journal of Chronic Obstructive Pulmonary Disease; 2012. **9**(1): p. 46-57. doi: 10.3109/15412555.2011.634863.
37. Ford, ES, JB Croft, DM Mannino, AG Wheaton, X Zhang and WH Giles, COPD Surveillance —United States, 1999-2011. Chest; 2013. **144**(1): p. 284-305. doi: 10.1378/chest.13-0809.
38. Tilert, T, C Dillon, R Paulose-Ram, E Hnizdo and B Doney, 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. Respiratory Research; 2013. **14**(1): p. 103. doi: 10.1186/1465-9921-14-103.
39. Zhang, X, JB Holt, H Lu, AG Wheaton, ES Ford, KJ Greenlund and JB Croft, 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. American Journal of Epidemiology; 2014. **179**(8): p. 1025-1033. doi: 10.1093/aje/kwu018.

40. Wheaton, AG, TJ Cunningham, ES Ford and JB Croft, Employment and activity limitations among adults with chronic obstructive pulmonary disease--United States, 2013, in *MMWR Morb Mortal Wkly Rep*. 2015. p. 289-95.
41. Shiels, MS, P Chernyavskiy, WF Anderson, AF Best, EA Haozous, P Hartge, PS Rosenberg, D Thomas, ND Freedman and A Berrington de Gonzalez, Trends in premature mortality in the USA by sex, race, and ethnicity from 1999 to 2014: an analysis of death certificate data. Lancet; 2017. **389**(10073): p. 1043-1054. doi: 10.1016/s0140-6736(17)30187-3.
42. Wheaton, AG, Y Liu, JB Croft, B VanFrank, TL Croxton, A Punturieri, L Postow and KJ Greenlund, Chronic Obstructive Pulmonary Disease and Smoking Status - United States, 2017. MMWR Morb Mortal Wkly Rep; 2019. **68**(24): p. 533-538. doi: 10.15585/mmwr.mm6824a1.
43. Terzikhan, N, KM Verhamme, A Hofman, BH Stricker, GG Brusselle and L Lahousse, Prevalence and incidence of COPD in smokers and non-smokers: the Rotterdam Study. Eur J Epidemiol; 2016. **31**(8): p. 785-92. doi: 10.1007/s10654-016-0132-z.
44. Doo, JH, SM Kim, YJ Park, KH Kim, YH Oh, JS Kim and SM Park, 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 Pulmonary Medicine; 2023. **23**(1): p. 237. doi: 10.1186/s12890-023-02533-1.
45. Perelman, J, J Alves, T-K Pfoertner, I Moor, B Federico, MAG Kuipers, M Richter, A Rimpela, AE Kunst and V Lorant, The association between personal income and smoking among adolescents: a study in six European cities. Addiction; 2017. **112**(12): p. 2248-2256. doi: https://doi.org/10.1111/add.13930.
46. Association, AL, COPD Causes and Risk Factors. 2024.
47. Clausen, A, ER Christensen, PR Jakobsen, J Søndergaard, B Abrahamsen and KH Rubin, Digital solutions for decision support in general practice – a rapid review focused on systems developed for the universal healthcare setting in Denmark. BMC Primary Care; 2023. **24**(1): p. 276. doi: 10.1186/s12875-023-02234-y.
48. Eriksen, J, M Ebbesen, KT Eriksen, C Hjermitslev, C Knudsen, P Bertelsen, C Nøhr and D Weber, Equity in digital healthcare - the case of Denmark. Front Public Health; 2023. **11**: p. 1225222. doi: 10.3389/fpubh.2023.1225222.
49. Joo, JY and MF Liu, Culturally tailored interventions for ethnic minorities: A scoping review. Nurs Open; 2021. **8**(5): p. 2078-2090. doi: 10.1002/nop2.733.
50. Radu, I, M Scheermesser, MR Spiess, C Schulze, D Händler-Schuster and J Pehlke-Milde, Digital Health for Migrants, Ethnic and Cultural Minorities and the Role of Participatory Development: A Scoping Review. Int J Environ Res Public Health; 2023. **20**(20). doi: 10.3390/ijerph20206962.
51. Sood, A, H Petersen, C Liu, O Myers, XW Shore, BA Gore, R Vazquez-Guillamet, LS Cook, P Meek and Y Tesfaigzi, Racial and Ethnic Minorities Have a Lower Prevalence of Airflow Obstruction than Non-Hispanic Whites. COPD: Journal of Chronic Obstructive Pulmonary Disease; 2022. **19**(1): p. 61-68. doi: 10.1080/15412555.2022.2029384.
52. Gilkes, A, S Hull, S Durbaba, P Schofield, M Ashworth, R Mathur and P White, Ethnic differences in smoking intensity and COPD risk: an observational study in primary care. NPJ Prim Care Respir Med; 2017. **27**(1): p. 50. doi: 10.1038/s41533-017-0052-8.
53. Hall, WJ, MV Chapman, KM Lee, YM Merino, TW Thomas, BK Payne, E Eng, SH Day and T Coyne-Beasley, Implicit Racial/Ethnic Bias Among Health Care Professionals and Its Influence on Health Care Outcomes: A Systematic Review. Am J Public Health; 2015.

105(12): p. e60-76. doi: 10.2105/ajph.2015.302903.

54. Dovidio, JF, LA Penner, TL Albrecht, WE Norton, SL Gaertner and JN Shelton, Disparities and distrust: The implications of psychological processes for understanding racial disparities in health and health care. Social Science & Medicine; 2008. **67**: p. 478–486. doi: 10.1016/j.socscimed.2008.03.019.
55. Williams, DR and M Sternthal, Understanding Racial-ethnic Disparities in Health: Sociological Contributions. Journal of Health and Social Behavior; 2010. **51**(1_suppl): p. S15-S27. doi: 10.1177/0022146510383838.
56. Tucker, CM, M Marsiske, KG Rice, JJ Nielson and K Herman, Patient-centered culturally sensitive health care: Model testing and refinement. Health Psychology; 2011. **30**: p. 342-350. doi: 10.1037/a0022967.
57. Chin, JL, Culturally competent health care. Public Health Reports; 2000. **115**(1): p. 25-33. doi: 10.1093/phr/115.1.25.
58. Berkowicz, R, F Palmgren, O Hertel and E Vignati, Using measurements of air pollution in streets for evaluation of urban air quality — meterological analysis and model calculations. Science of The Total Environment; 1996. **189-190**: p. 259-265. doi: 10.1016/0048-9697(96)05217-5.
59. Vardoulakis, S, BEA Fisher, K Pericleous and N Gonzalez-Flesca, Modelling air quality in street canyons: a review. Atmospheric Environment; 2003. **37**(2): p. 155-182. doi: 10.1016/S1352-2310(02)00857-9.
60. Jonidi Jafari, A, E Charkhloo and H Pasalari, Urban air pollution control policies and strategies: a systematic review. J Environ Health Sci Eng; 2021. **19**(2): p. 1911-1940. doi: 10.1007/s40201-021-00744-4.
61. Ward, F, HJ Lowther-Payne, EC Halliday, K Dooley, N Joseph, R Livesey, P Moran, S Kirby and J Cloke, Engaging communities in addressing air quality: a scoping review. Environ Health; 2022. **21**(1): p. 89. doi: 10.1186/s12940-022-00896-2.
62. Rosen, LJ, V Myers, JP Winickoff and J Kott, Effectiveness of Interventions to Reduce Tobacco Smoke Pollution in Homes: A Systematic Review and Meta-Analysis. Int J Environ Res Public Health; 2015. **12**(12): p. 16043-59. doi: 10.3390/ijerph121215038.
63. Titus, AR, L Kalousova, R Meza, DT Levy, JF Thrasher, MR Elliott, PM Lantz and NL Fleischer, Smoke-Free Policies and Smoking Cessation in the United States, 2003-2015. Int J Environ Res Public Health; 2019. **16**(17). doi: 10.3390/ijerph16173200.
64. Elshawi, R, MH Al-Mallah and S Sakr, On the interpretability of machine learning-based model for predicting hypertension. BMC Medical Informatics and Decision Making; 2019. **19**(1): p. 146. doi: 10.1186/s12911-019-0874-0.
65. Hakkoum, H, I Abnane and A Idri, Interpretability in the medical field: A systematic mapping and review study. Applied Soft Computing; 2022. **117**: p. 108391. doi: https://doi.org/10.1016/j.asoc.2021.108391.
66. Genuer, R, J-M Poggi and C Tuleau-Malot, Variable selection using random forests. Pattern recognition letters; 2010. **31**(14): p. 2225-2236. doi: 10.1016/j.patrec.2010.03.014.
67. Grömping, U, Variable importance assessment in regression: linear regression versus random forest. The American Statistician; 2009. **63**(4): p. 308-319. doi: 10.1198/tast.2009.08199.
68. Adadi, A and M Berrada, Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access 2018. **6**: p. 52138-52160. doi: 10.1109/ACCESS.2018.2870052.
69. Lundberg, S and S-I Lee. A Unified Approach to Interpreting Model Predictions. in *NIPS*. 2017. Long Beach, CA.
70. Kamath, SS and VS Ananthanarayana, Semantics-based Web service classification using morphological analysis and ensemble learning techniques. International Journal of Data Science and Analytics; 2016. **2**: p. 61–74. doi: 10.1007/s41060-016-0026-x.

71. Gilpin, LH, D Bau, BZ Yuan, A Bajwa, M Specter and L Kagal, Explaining Explanations: An Overview of Interpretability of Machine Learning. arXiv:1806.00069 [cs.AI]; 2018. doi: 10.1109/DSAA.2018.00018.
72. Kottemann, JE and WE Remus, A Study of the Relationship Between Decision Model Naturalness and Performance. MIS Quarterly; 1989. **13**(2): p. 170-181. doi: 10.2307/248924.
73. Mannino, DM. Chronic Obstructive Pulmonary Disease: Definition and Epidemiology. in *Respiratory Care*. 2003.
74. Black, C, Y Tesfaigzi, JA Bassein and LA Miller, Wildfire smoke exposure and human health: Significant gaps in research for a growing public health issue. Environmental Toxicology and Pharmacology; 2017. **55**: p. 186-195. doi: 10.1016/j.etap.2017.08.022.
75. Kelley, T and GD Kearney, Insights Into the Environmental Health Burden of Childhood Asthma. Environmental Health Insights; 2018. **12**: p. 117863021875744. doi: 10.1177/1178630218757445.
76. Wells, WD, Life Style and Psychographics, Chapter 13: Life Style and Psychographics: Definitions, Uses, and Problems. 2011: Marketing Classics Press.
77. Assael, H, A demographic and psychographic profile of heavy internet users and users by type of internet usage. Journal of Advertising Research; 2005. **45**(1): p. 93-123. doi: 10.1017/ S0021849905050014
78. Fox, S and K Purcell, Chronic disease and the internet. 2010: Pew Internet & American Life Project Washington, DC.
79. Sujaritpong, S, K Dear, M Cope, S Walsh and T Kjellstrom, Quantifying the health impacts of air pollution under a changing climate—a review of approaches and methodology. International Journal of Biometeorology; 2014. **58**(2): p. 149-160. doi: 10.1007/s00484-012-0625-8.
80. Ramanathan, V and Y Feng, Air pollution, greenhouse gases and climate change: Global and regional perspectives. Atmospheric Environment; 2009. **43**(1): p. 37-50. doi: 10.1016/j.atmosenv.2008.09.063.
81. Barbour, E and EA Deakin, Smart Growth Planning for Climate Protection. Journal of the American Planning Association; 2012. **78**(1): p. 70-86. doi: 10.1080/01944363.2011.645272.
82. Boyce, JK and M Pastor, Clearing the air: incorporating air quality and environmental justice into climate policy. Climatic change; 2013. **120**(4): p. 801-814. doi: 10.1007/s10584-013-0832-2.
83. Cushing, L, D Blaustein-Rejto, M Wander, M Pastor, J Sadd, A Zhu and R Morello-Frosch, Carbon trading, co-pollutants, and environmental equity: Evidence from California’s cap-and-trade program (2011–2015). PLoS Medicine; 2018. **15**(7): p. e1002604. doi: 10.1371/journal.pmed.1002604.
84. Kinney, PL, Climate change, air quality, and human health. American Journal of Preventive Medicine; 2008. **35**(5): p. 459-467. doi: 10.1016/j.amepre.2008.08.025.
85. Strickland, E, Andrew Ng: Unbiggen AI The AI pioneer says it’s time for smart-sized, “data-centric” solutions to big issues, in *IEEE Spectrum*. 2022.
86. Liu, JC, G Pereira, SA Uhl, MA Bravo and ML Bella, A systematic review of the physical health impacts from non-occupational exposure to wildfire smoke. Environmental Research; 2015. **136**. doi: 10.1016/j.envres.2014.10.015.
87. Kamis, A, R Cao, Y He, Y Tian and C Wu, Predicting Lung Cancer in the United States: A Multiple Model Examination of Public Health Factors. International Journal of Environmental Research and Public Health; 2021. **18**(11). doi: 10.3390/ijerph18116127.
88. Andre, M, K Sartelet, S Moukhtar, JM Andre and M Redaelli, Diesel, petrol or electric vehicles: What choices to improve urban air quality in the Ile-de-France region? A simulation platform and case study. Atmospheric Environment; 2020. **241**: p. 117752. doi: 10.1016/j.atmosenv.2020.117752.

89. Lam, YF, JS Fu, S Wu and LJ Mickley, Impacts of future climate change and effects of biogenic emissions on surface ozone and particulate matter concentrations in the United States. Atmospheric Chemistry and Physics; 2011. **11**(10): p. 4789-4806. doi: 10.5194/acp-11-4789-2011.

90. Anenberg, SC, LW Horowitz, DQ Tong and JJ West, An Estimate of the Global Burden of Anthropogenic Ozone and Fine Particulate Matter on Premature Human Mortality Using Atmospheric Modeling. Environmental Health Perspectives; 2010. **118**(9): p. 1189-1195. doi: doi:10.1289/ehp.0901220.

91. Pope, CA, ML Hansen, RW Long, KR Nielsen, NL Eatough, WE Wilson and DJ Eatough, Ambient particulate air pollution, heart rate variability, and blood markers of inflammation in a panel of elderly subjects. Environmental Health Perspectives; 2004. **112**(3). doi: 10.1289/ehp.6588.

92. Araujo, JA, Particulate air pollution, systemic oxidative stress, inflammation, and atherosclerosis. Air Quality, Atmosphere & Health; 2011. **4**: p. 79–93. doi: 10.1007/s11869-010-0101-8.

93. Dockery, DW, CA Pope, X Xu, JD Spengler, JH Ware, ME Fay, J Benjamin G. Ferris and FE Speizer, An Association between Air Pollution and Mortality in Six U.S. Cities. The New England Journal of Medicine; 1993. **329**(24): p. 1753-1759. doi: 10.1056/NEJM199312093292401.

94. Belleudi, V, A Faustini, M Stafoggia, G Cattani, A Marconi, CA Perucci and F Forastiere, Impact of fine and ultrafine particles on emergency hospital admissions for cardiac and respiratory diseases. Epidemiology; 2010: p. 414-423. doi: 10.1097/EDE.0b013e3181d5c021.

95. Schraufnagel, DE, The health effects of ultrafine particles. Experimental & molecular medicine; 2020. **52**(3): p. 311-317. doi: 10.1038/s12276-020-0403-3.

96. Zhang, Y, Z Ding, Q Xiang, W Wang, L Huang and F Mao, Short-term effects of ambient PM1 and PM2.5 air pollution on hospital admission for respiratory diseases: case-crossover evidence from Shenzhen, China. International journal of hygiene and environmental health; 2020. **224**: p. 113418. doi: 10.1016/j.ijheh.2019.11.001.

97. Ni, L, C-C Chuang and L Zuo, Fine particulate matter in acute exacerbation of COPD. Frontiers in physiology; 2015. **6**: p. 294. doi: 10.3389/fphys.2015.00294.

98. Li, T, R Hu, Z Chen, Q Li, S Huang, Z Zhu and L-F Zhou, Fine particulate matter (PM2. 5): The culprit for chronic lung diseases in China. Chronic diseases and translational medicine; 2018. **4**(03): p. 176-186. doi: 10.1016/j.cdtm.2018.07.002.

99. Kim, CS, NE Alexis, AG Rappold, H Kehrl, MJ Hazucha, JC Lay, MT Schmitt, M Case, RB Devlin and DB Peden, Lung function and inflammatory responses in healthy young adults exposed to 0.06 ppm ozone for 6.6 hours. American journal of respiratory and critical care medicine; 2011. **183**(9): p. 1215-1221. doi: 10.1164/rccm.201011-1813OC.

100. Mudway, IS and FJ Kelly, Ozone and the lung: a sensitive issue. Molecular aspects of medicine; 2000. **21**(1-2): p. 1-48. doi: 10.1016/S0098-2997(00)00003-0.

101. Uysal, N and RM Schapira. Effects of ozone on lung function and lung diseases. Current opinion in pulmonary medicine 2003.

102. Hendryx, M, J Luo, C Chojenta and JE Byles, Air pollution exposures from multiple point sources and risk of incident chronic obstructive pulmonary disease (COPD) and asthma. Environmental research; 2019. **179**: p. 108783. doi: 10.1016/j.envres.2019.108783.

103. Chen, T-M, WG Kuschner, J Gokhale and S Shofer, Outdoor air pollution: nitrogen dioxide, sulfur dioxide, and carbon monoxide health effects. The American Journal of the Medical Sciences; 2007. **333**(4): p. 249-256. doi: 10.1097/MAJ.0b013e31803b900f.

104. Gong, H, Jr., WS Linn, KW Clark, KR Anderson, MD Geller and C Sioutas, Respiratory responses to exposures with fine particulates and nitrogen dioxide in the elderly with and without COPD. Inhalation Toxicology; 2005. **17**(3): p. 123-132. doi:

10.1080/08958370590904481.

105. Lary, DJ, T Lary and B Sattler, Using Machine Learning to Estimate Global PM2.5 for Environmental Health Studies. Environmental Health Insights; 2015. **9s1**: p. EHI.S15664. doi: 10.4137/ehi.s15664.

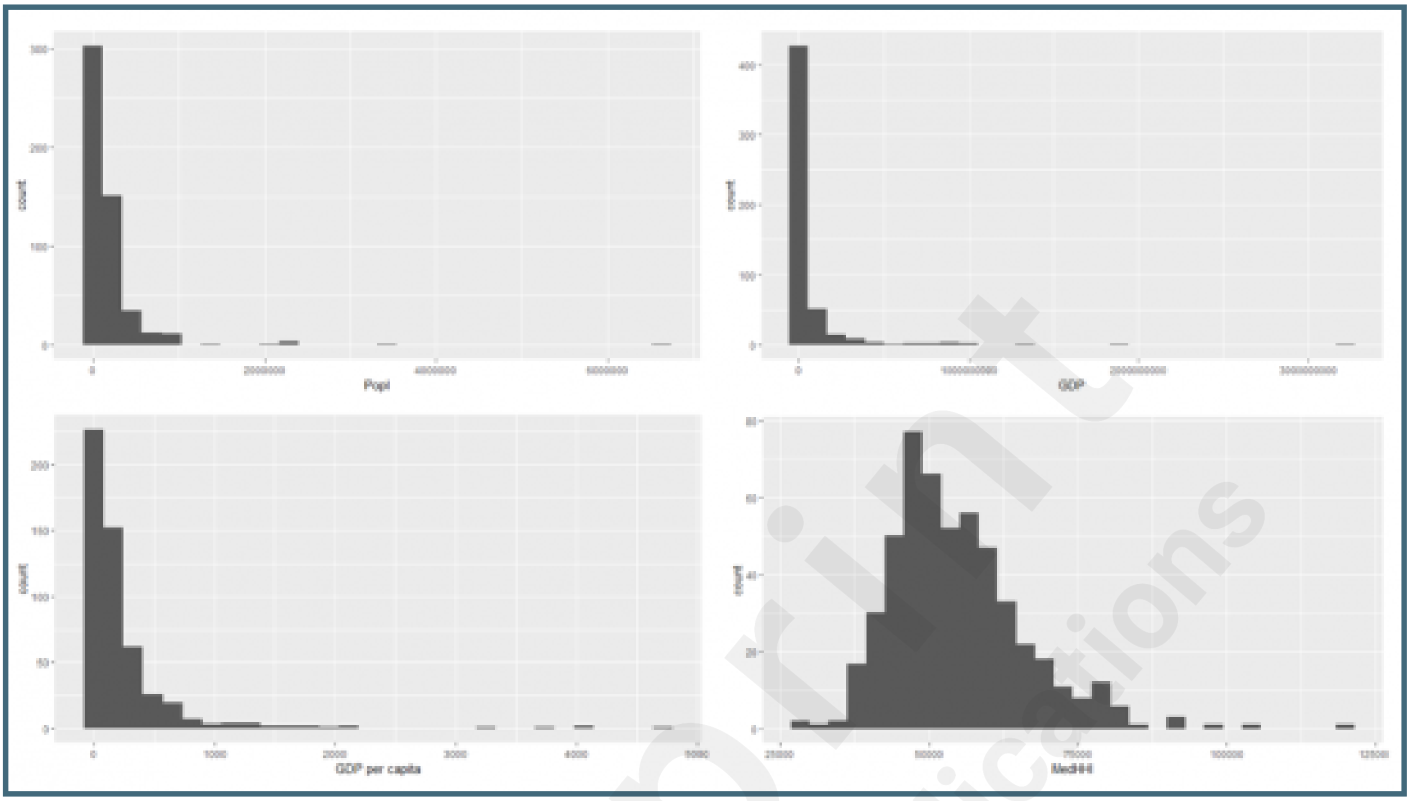
106. Tai, APK, LJ Mickley, DJ Jacob, EM Leibensperger, L Zhang, JA Fisher and HOT Pye, Meteorological modes of variability for fine particulate matter (PM2.5) air quality in the United States: implications for PM2.5 sensitivity to climate change. Atmos. Chem. Phys.; 2012. **12**(6): p. 3131-3145. doi: 10.5194/acp-12-3131-2012.

107. Peterson, GCL, C Hogrefe, AE Corrigan, LM Neas, R Mathur and AG Rappold, Impact of reductions in emissions from major source sectors on fine particulate matter–related cardiovascular mortality. Environmental Health Perspectives; 2020. **128**(1): p. 017005. doi: 10.1289/EHP5692.

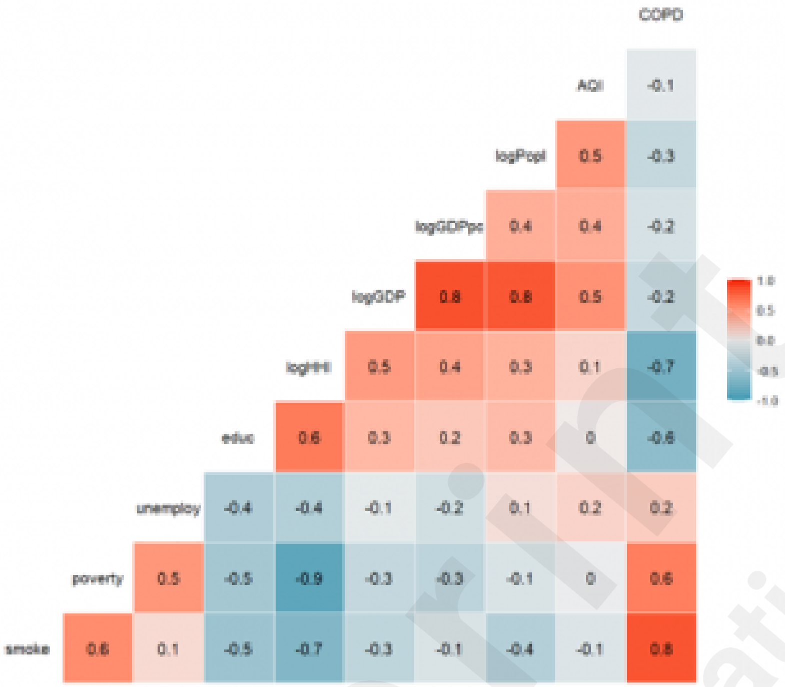
Supplementary Files

Figures

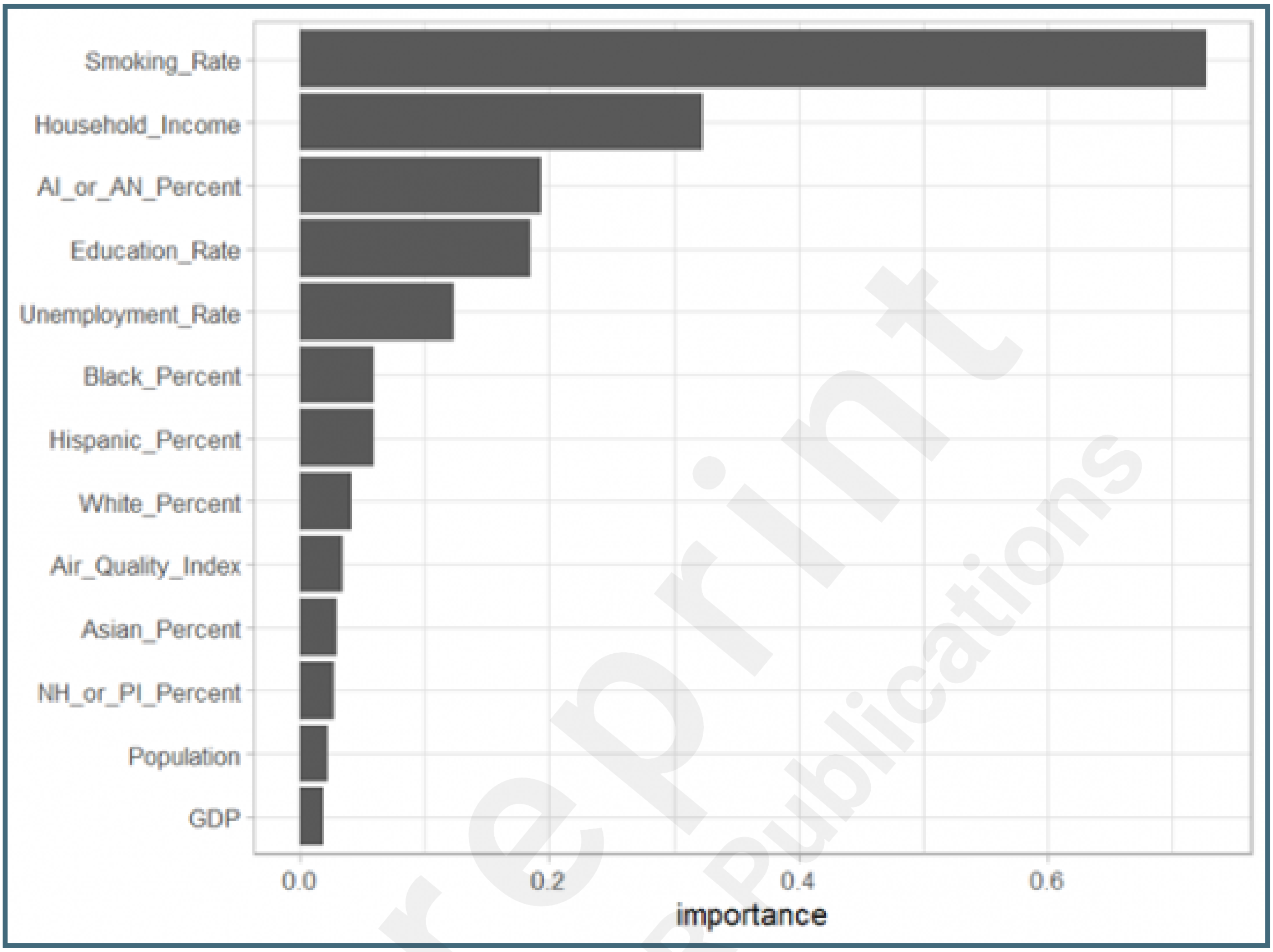
Population, GDP, GDP per capita, and median household income.



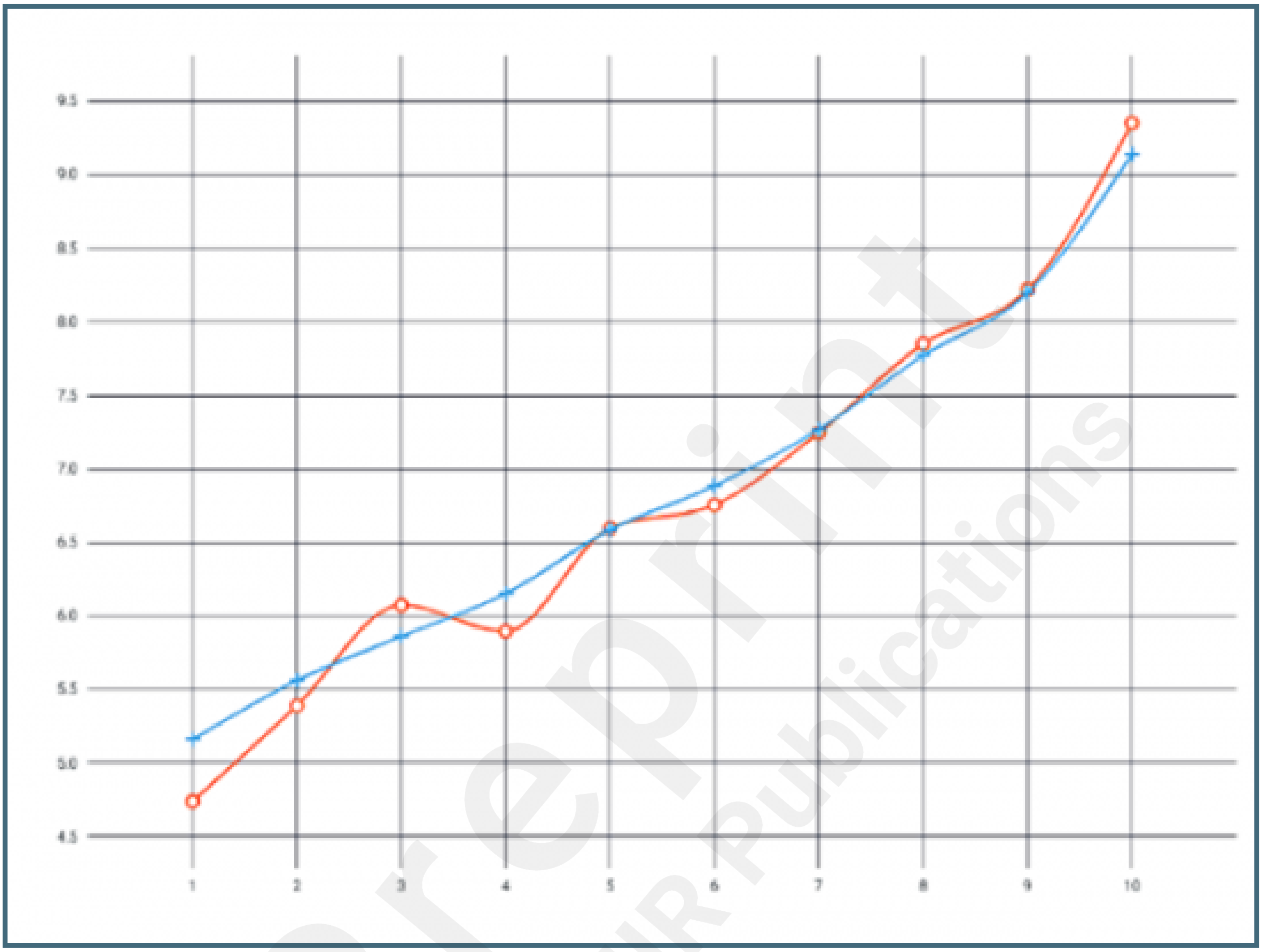
Correlations among main variables.



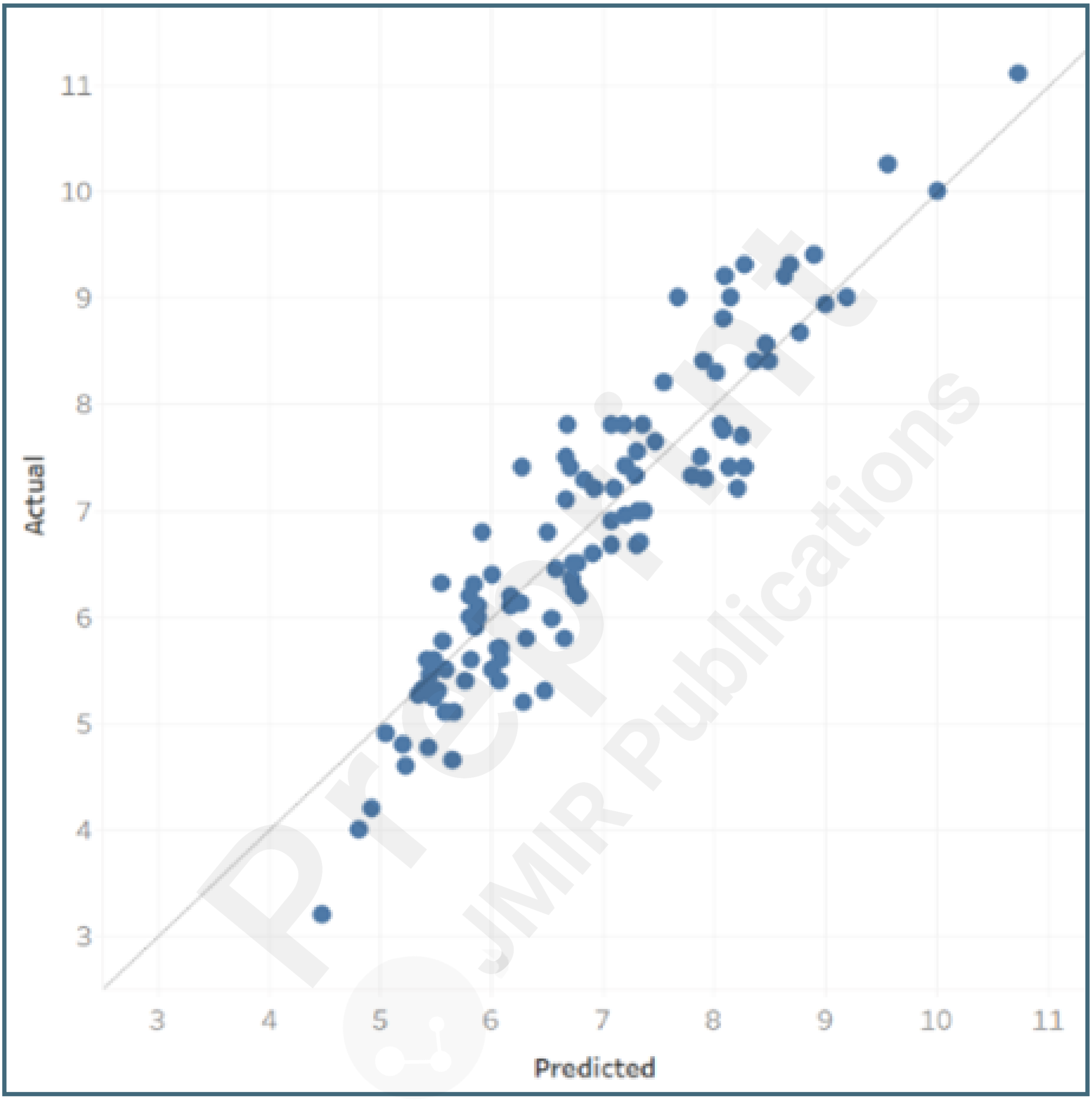
Importance of Model Variables for Gradient Boosted Tree.



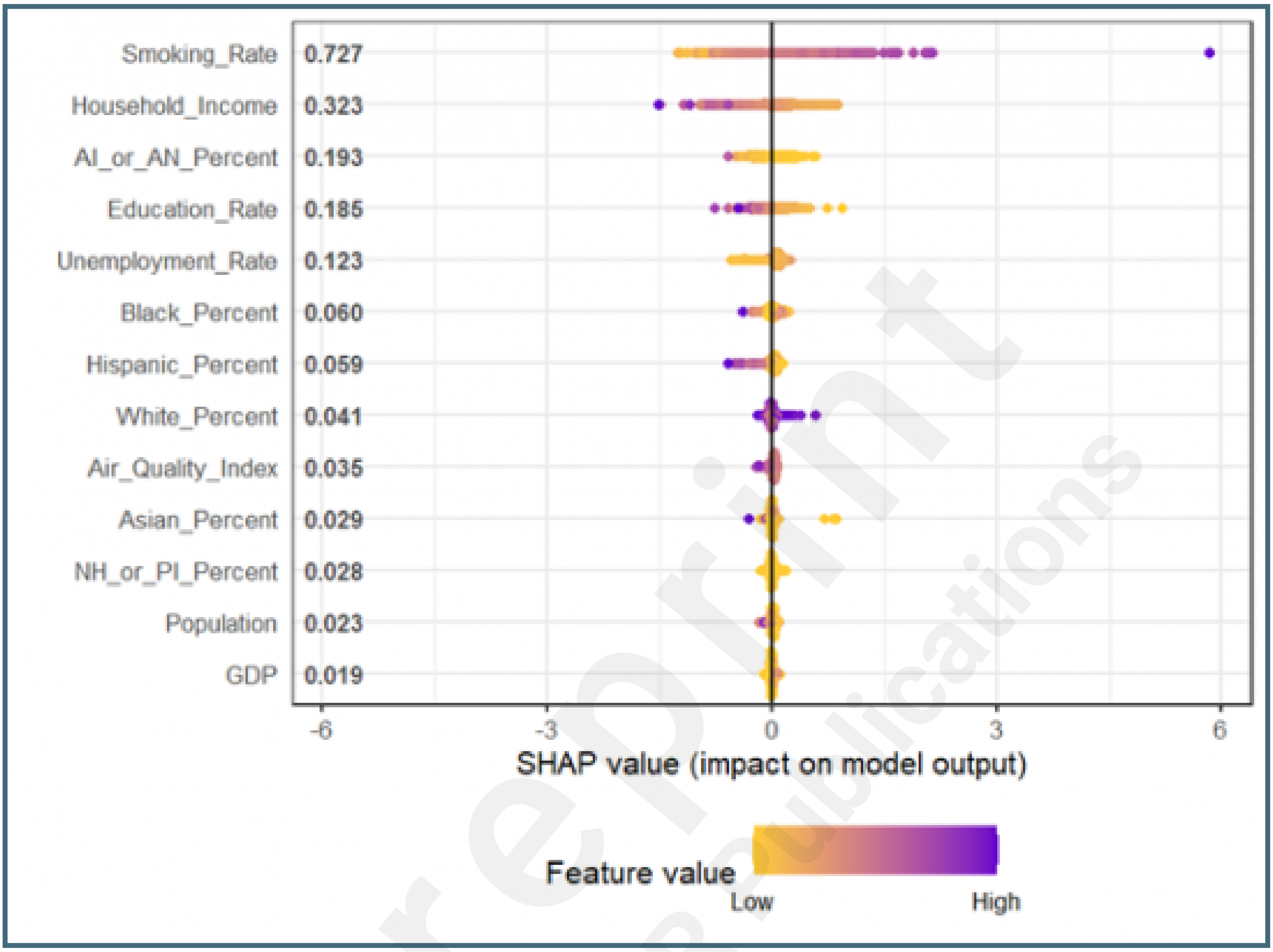
Lift plot showing COPD rate as a function of 10 decile bins. Predicted Values (blue), Actual Values (red).



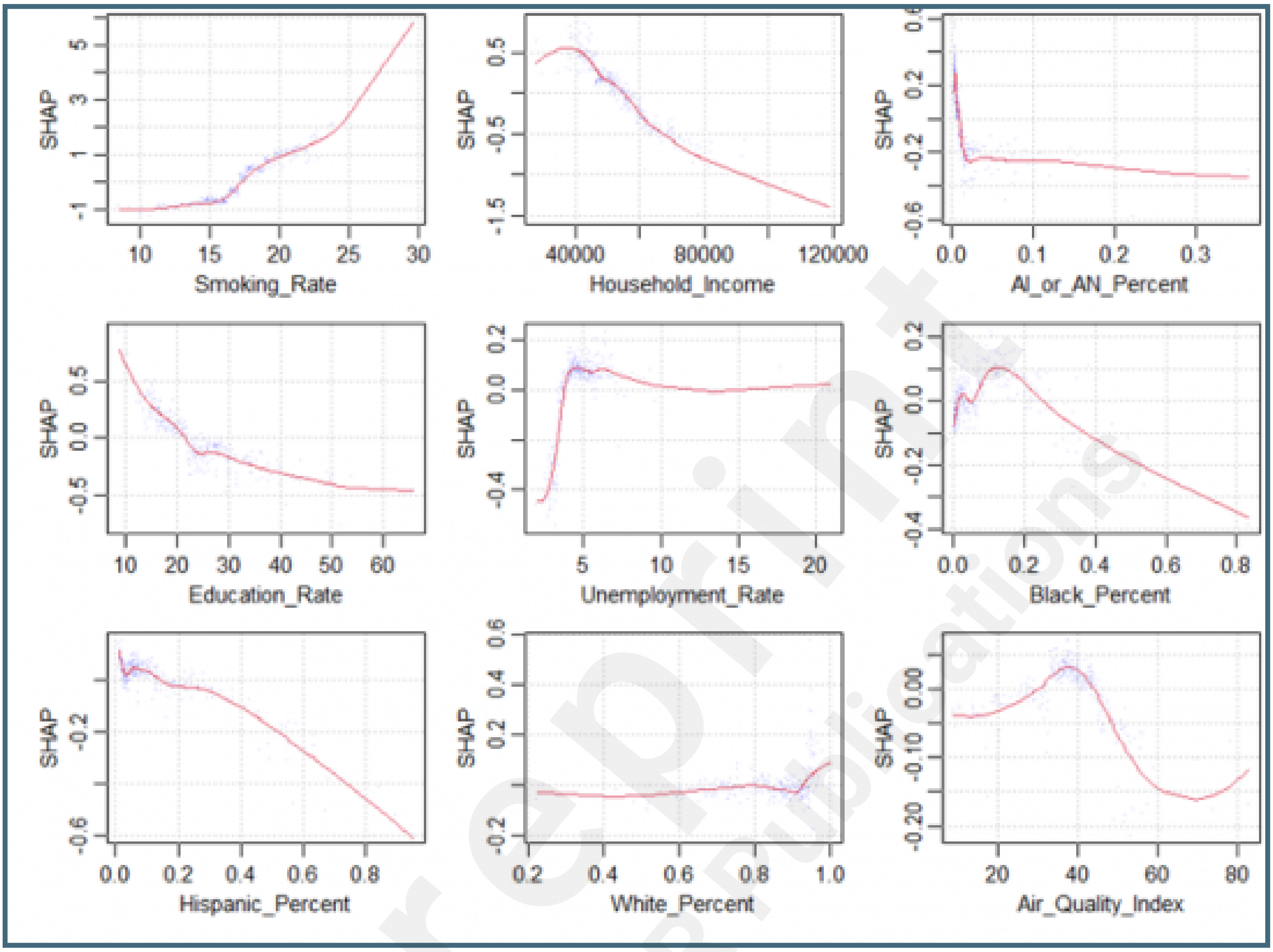
Prediction residuals.



SHAP Values for All Features (Variables).



SHAP Plots for the nine most important variables.



TOC/Feature image for homepages

Photograph of lung anatomy. Credit: Robina Weermeijer.