A Bayesian Hierarchical Approach to Estimating PM2.5 Effects on COPD Prevalence Across the United States

Hayato Isa

University of California, Berkeley

Abstract

This study investigates the relationship between state-level adult COPD (Chronic Obstructive Pulmonary Disease) prevalence and key predictors, smoking rates, diabetes prevalence, poverty rates, and fine particulate matter ($PM_{2.5}$) concentration, in the United States. To account for persistent state-specific effects and quantify uncertainty, we use a Bayesian hierarchical model. Posterior inference indicates positive associations between COPD prevalence and smoking rates (posterior mean $\mathbb{E}[\beta_1] = 0.165$, 94% HDI [0.083, 0.253]), diabetes prevalence ($\mathbb{E}[\beta_2] = 0.460$, [0.359, 0.559]), and poverty rates ($\mathbb{E}[\beta_3] = 0.066$, [-0.025, 0.150]). The model suggests a nonlinear $PM_{2.5}$ effect, with a negative linear coefficient ($\mathbb{E}[\beta_4] = -0.449$) and a positive quadratic coefficient ($\mathbb{E}[\beta_5] = 0.026$). State-specific effects are substantial and may reflect regional factors such as historical coal mining and ethnic composition. In a predictive task for 2020 COPD prevalence, the hierarchical model (RMSE = 0.5617) outperforms a naive linear regression extrapolation (RMSE = 0.7767), showing the predictive advantage of integrating information from multiple sources. While the analysis establishes robust associations and shows the existence of state-level diversity, it cannot support direct causal claims. Results suggest that public health strategies targeting COPD should consider region-specific risk profiles, including air quality, socioeconomic conditions, and historical environmental legacies.

Keywords— COPD, Bayesian hierarchical model, PM2.5, state-level health disparities, public health

Contents

1 Data Overview

Data	Description
Y: COPD_rate [1]	State-level annual COPD prevalence values, constructed by filtering BRFSS records reporting overall crude prevalence of Chronic Obstructive Pulmonary Disease among adults aged 18 and older. (Based on a population sample collected through BRFSS random telephone surveys)
X_1: Smoking_rate [2]	State-level annual smoking prevalence values, constructed by filtering BRFSS responses indicating "Yes" to current smoker status to obtain overall crude prevalence for each state and year. (Based on a population sample collected through BRFSS random telephone surveys)
X_2: Diabetes_rate [2]	State-level annual diabetes prevalence values, constructed by filtering BRFSS responses indicating "Yes" to the question, "Have you ever been told by a doctor that you have diabetes?", yielding overall crude prevalence. (Based on a population sample collected through BRFSS random telephone surveys)
X_3: Poverty_rate [3]	State-level annual poverty rates based on the U.S. Census Bureau's official poverty thresholds; no additional filtering was required. (Based on the American Community Survey, a population sample provided by the U.S. Census Bureau)
X_4: PM2.5_level [4]	State-level annual $PM_{2.5}$ concentration values, constructed by averaging county-level $PM_{2.5}$ measurements within each state for every year. (Based on census-tract–level measurements provided by the National Environmental Public Health Tracking Network)

Since BRFSS relies on telephone-based surveys, households without telephone access (particularly those from lower-income populations) may be underrepresented. Similarly, the U.S. Census Bureau faces potential sampling bias due to nonresponse. To mitigate these biases, both BRFSS and the U.S. Census Bureau apply post-stratification weighting procedures to reduce sampling bias [5][6]. Nevertheless, measurement error is possible for self-reported variables such as smoking status and diabetes diagnosis, which may be influenced by recall bias or social desirability bias. PM2.5 concentrations are estimated from environmental monitoring data to produce census tract–level estimates across all the counties in 48 states, excluding Alaska and Hawaii.

All health and socioeconomic variables are reported at the state–year level, whereas PM2.5 data are originally available at the county and daily level. To ensure consistency in the unit of analysis, PM2.5 measurements were aggregated to state–year averages. This level of aggregation enables the identification of broad spatial and temporal patterns. More granular stratifications, such as prevalence estimates by ethnicity, would enable a deeper examination of health disparities. However, such information was inconsistently reported across datasets. Missing values primarily reflected observations that failed to meet data providers' reliability thresholds. Because missingness was minimal, state–year observations with missing values were simply dropped from the analysis.

2 Research Question

Question: How do PM2.5 concentration and other factors relate to adult COPD prevalence at the state level in the United States?

Answering this research question can inform practical decisions such as determining the priority of air pollution (PM2.5) reduction, and designing region-specific public health policies and resource allocation.

This study uses a Bayesian hierarchical model to analyze the relationship between state-level COPD prevalence and multiple predictors. It explicitly models state-specific effects, and uncertainty is expressed probabilistically, enabling risk-aware decision-making.

However, we should note that the method estimates associations, not causal effects, so conclusions about causality cannot be drawn directly.

3 Prior Work

The study [7] examines the prevalence and incidence of chronic obstructive pulmonary disease (COPD) among smokers and non-smokers using data from the Rotterdam Study, highlighting a positive association between smoking and COPD prevalence. Their dataset includes detailed information on age and sex, allowing analysis within specific subgroups. In contrast, our study only uses overall smoking rates and COPD prevalence, preventing us from exploring these subgroup-level relationships.

Another study [14] applies a multivariable mixed linear model and a time-varying Cox model to assess the effects of changes in PM2.5 on annual lung function decline and COPD incidence, respectively, showing a positive association between PM2.5 exposure and COPD risk. Our simpler Bayesian hierarchical model does not account for temporal variation, which may limit the accuracy of our findings.

4 EDA

As shown in the figure below, states with low COPD prevalence tend to remain low over time, whereas states with higher prevalence remain consistently high. This persistence suggests the presence of state-specific structural or environmental conditions that influence COPD risk, which supports the use of a Bayesian hierarchical model.

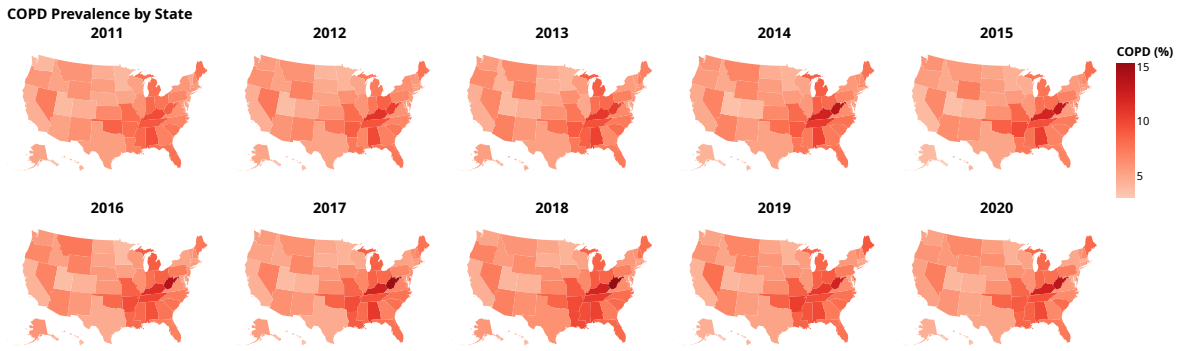
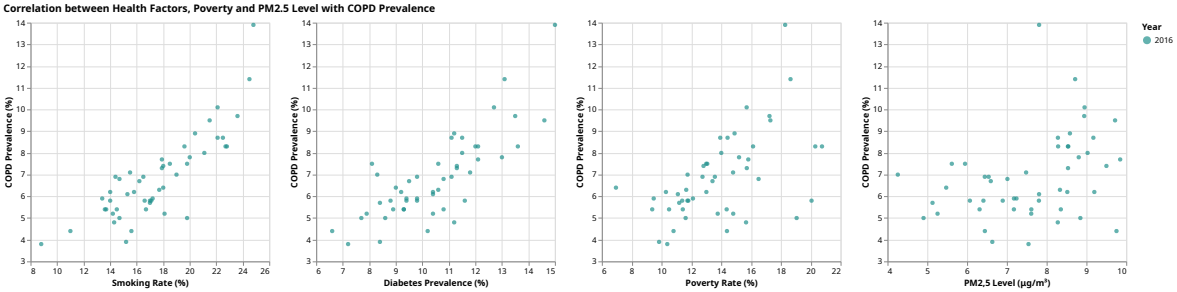
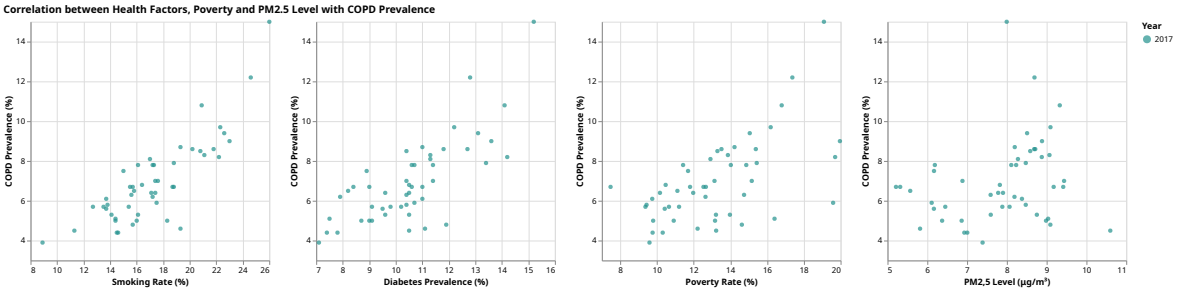
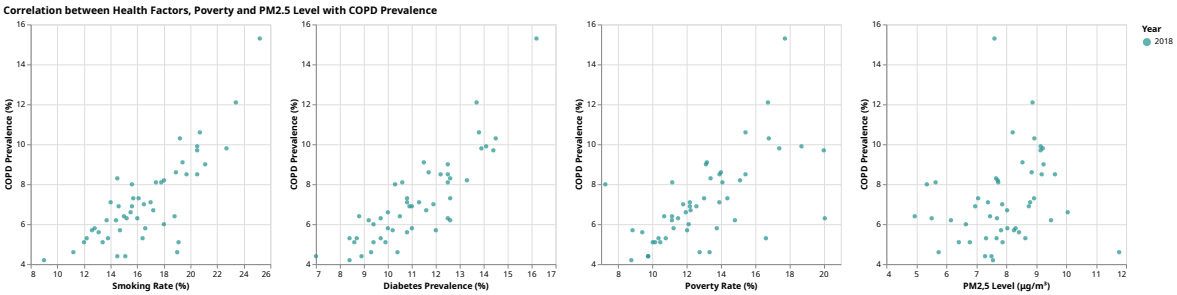


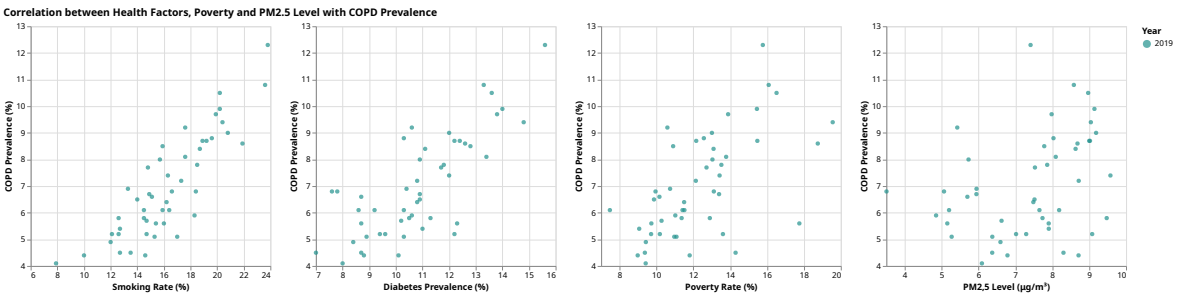
Figure 1: State-level COPD prevalence from 2011–2020

Furthermore, for each year, COPD prevalence shows a roughly linear association with smoking rate, diabetes prevalence, and poverty rate, and a quadratic association with PM2.5 levels. These observed patterns justify including smoking rate, diabetes prevalence, poverty rate, and PM2.5 as covariates in our model to predict COPD prevalence.


Correlation between Health Factors, Poverty and PM2.5 Level with COPD Prevalence
COPD Prevalence (%)
Smoking Rate (%)
Diabetes Prevalence (%)
Poverty Rate (%)
PM2,5 Level (µg/m³)
Year
2016


Correlation between Health Factors, Poverty and PM2.5 Level with COPD Prevalence
COPD Prevalence (%)
Smoking Rate (%)
Diabetes Prevalence (%)
Poverty Rate (%)
PM2,5 Level (µg/m³)
Year
2017


Correlation between Health Factors, Poverty and PM2.5 Level with COPD Prevalence
COPD Prevalence (%)
Smoking Rate (%)
Diabetes Prevalence (%)
Poverty Rate (%)
PM2,5 Level (µg/m³)
Year
2018


Correlation between Health Factors, Poverty and PM2.5 Level with COPD Prevalence
COPD Prevalence (%)
Smoking Rate (%)
Diabetes Prevalence (%)
Poverty Rate (%)
PM2,5 Level (µg/m³)
Year
2019

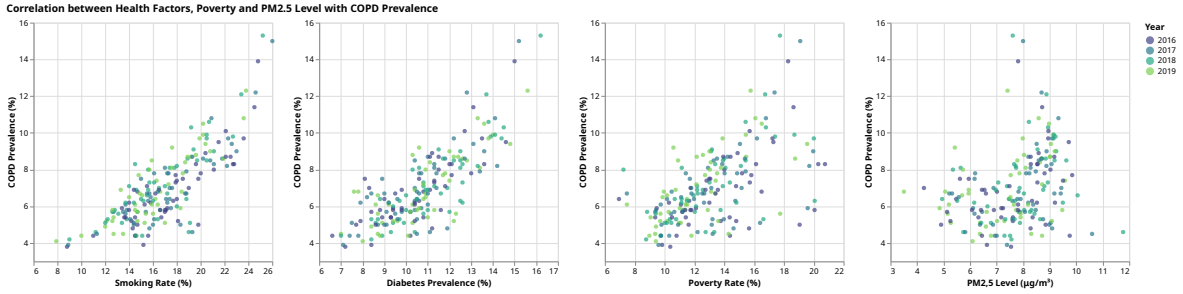


Figure 2: Correlation between COPD prevalence and state-level predictors

5 How do PM2.5 concentration and other factors relate to adult COPD prevalence at the state level in the United States?: Posterior Inference and Analysis in a Bayesian Hierarchical Model

5.0.1 Methods

The risk of developing Chronic Obstructive Pulmonary Disease (COPD) is substantially higher among adults with a history of cigarette smoking across both younger and older populations [7, 8]. This association is biologically well-established. Cigarette smoke generates large amounts of reactive oxygen species, which induce oxidative stress, airway inflammation, and structural lung damage [9]. Obesity (linked to diabetes) is also a risk factor for COPD. Long-term obesity and progressive weight gain in adulthood are associated with higher COPD incidence compared with individuals who maintain a normal weight [10]. Mechanically, dysfunctional adipocytes contribute to systemic inflammation and promote pro-inflammatory signaling within the lung microenvironment [11].

Socioeconomic disadvantage further increases COPD risk. Individuals living below the relative poverty line exhibit higher COPD prevalence, particularly among older adults [12]. Poverty influences respiratory health through multiple pathways, including prenatal exposures, increased frequency of childhood respiratory infections, substandard housing conditions, greater exposure to indoor and outdoor air pollution, and limited access to preventive healthcare [13]. In addition, long-term exposure to fine particulate matter (PM2.5) contributes to the decline of lung function and increased COPD risk [14], as PM2.5 can penetrate deeply into the pulmonary system and induce direct epithelial and inflammatory injury [15].

Motivated by these established epidemiological pathways, our model incorporates state-level smoking prevalence, diabetes rate, poverty rate, and mean PM2.5 concentration for each year as predictors of annual state-level COPD prevalence.

Next, as shown in Figure 3, state-level COPD prevalence also exhibits strong temporal stability. States with low COPD prevalence tend to remain low over time, whereas states with higher prevalence remain consistently high. This persistence suggests the presence of state-specific structural or environmental conditions that influence COPD risk other than those covariates. To account for this assumption, we introduce a state-specific latent effect u_i, which captures persistent deviations in COPD prevalence due to underlying state-specific factors.

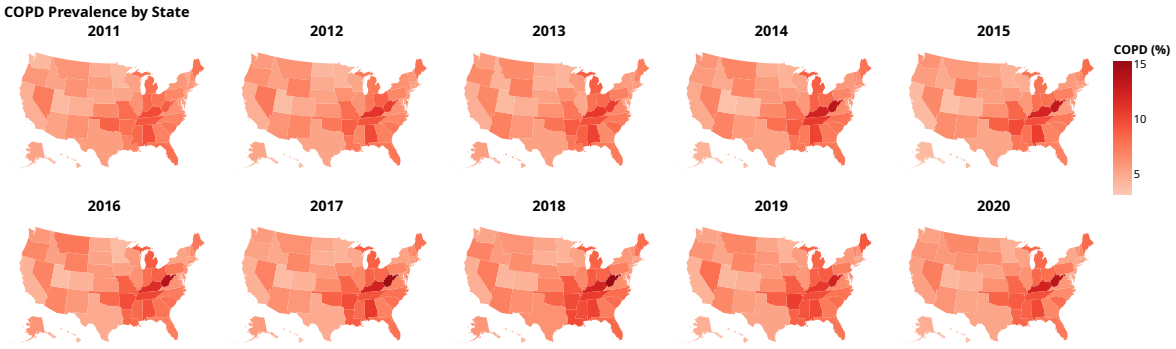


Figure 3: State-level COPD prevalence from 2011–2020

As shown in Figure 3, state-level COPD prevalence typically falls within a relatively narrow range (approximately 5–15%). Thus, the magnitude of both the state-specific effects u_i and the error term ϵ is expected to be modest. To encode this prior knowledge, we place weakly informative Half-Normal($\sigma = 1$) priors on the standard deviations σ_u and σ_Y.

Furthermore, the scatter plots in Figure 4 indicate that X_1, X_2 and X_3 exhibit approximately linear associations with Y (Pearson correlation coefficients $r = 0.812, 0.792, 0.576$, respectively), whereas X_4 displays a clear nonlinear (quadratic) relationship. Visual inspection of the fitted lines also indicates that the slopes of the linear associations are modest in magnitude. Incorporating this empirical knowledge, we specify the following Bayesian hierarchical model:

$$\sigma_u, \sigma_Y \sim \text{Half-Normal}(\sigma = 1), \quad u_i \sim \mathcal{N}(0, \sigma_u), \quad \epsilon \sim \mathcal{N}(0, \sigma_Y), \quad \beta_i \sim \mathcal{N}(0, 1)$$

$$Y_i = u_i + \beta_0 + \beta_1 X_1^i + \beta_2 X_2^i + \beta_3 X_3^i + \beta_4 X_4^i + \beta_5 (X_4^i)^2 + \epsilon$$

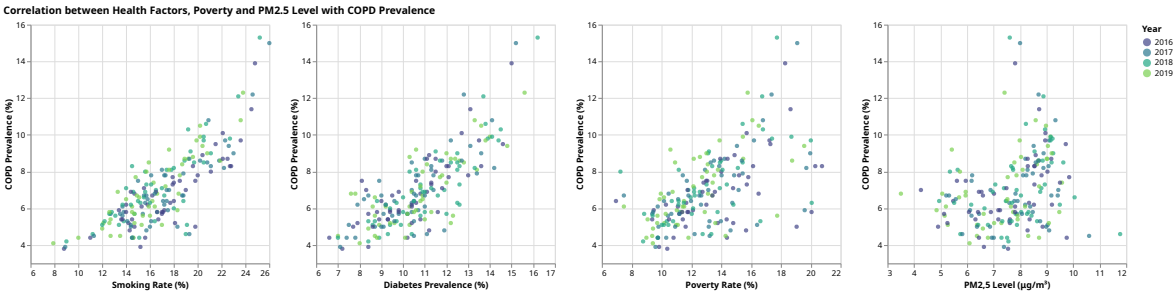


Figure 4: Correlation between COPD prevalence and state-level predictors

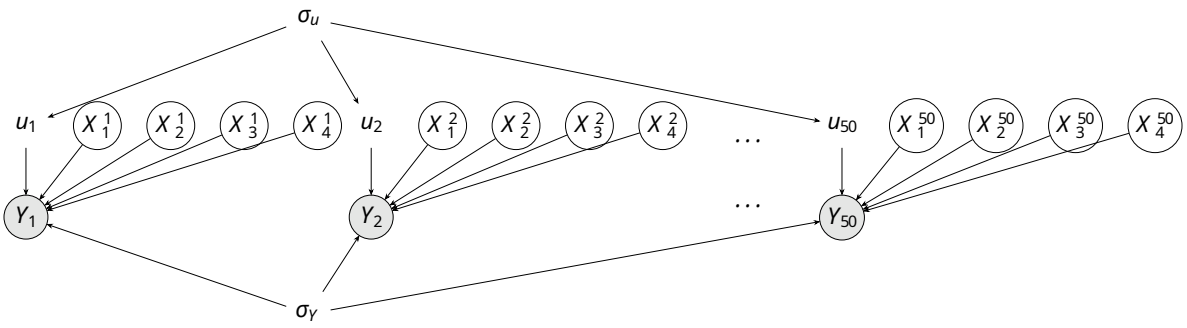


Figure 5: Bayesian hierarchical model for predicting state-level COPD prevalence

Table 1: Description of Model Parameters and Variables

Symbol	Description
σ_u	Standard deviation of the state-level latent effect.
σ_Y	Standard deviation of the observation-level residual variation in COPD prevalence.
u_i	State-specific latent effect on COPD prevalence in state i.
Y_i	Observed COPD prevalence for state i.
X_1^i	Smoking prevalence in state i.
X_2^i	Diabetes rate in state i.
X_3^i	Poverty rate in state i.
X_4^i	Annual mean $PM_{2.5}$ concentration in state i.

5.0.2 Results

As presented in Table 2, the baseline intercept is estimated as $\mathbb{E}[\beta_0] = 0.265$. The 94% Highest Density Interval (HDI) for β_1 is [0.083, 0.253], providing strong posterior evidence for a positive relationship between state-level smoking prevalence and COPD prevalence. The posterior mean $\mathbb{E}[\beta_1] = 0.165$ implies that, holding all other predictors constant, a one-percentage-point increase in state smoking prevalence is associated with an expected 0.165 percentage-point increase in state COPD prevalence.

Similarly, the 94% HDI for β_2 is [0.359, 0.559], providing strong posterior evidence for a positive relationship between state-level diabetes rate and COPD prevalence. The posterior mean $\mathbb{E}[\beta_2] = 0.460$ indicates that a one-percentage-point increase in state diabetes rate corresponds, on average, to a 0.460 percentage-point increase in state COPD prevalence, controlling for other covariates. For β_3, the posterior mean $\mathbb{E}[\beta_3] = 0.066$ suggests that, holding other factors constant, a one-percentage-point increase in the state poverty rate is associated with an expected 0.066 percentage-point increase in COPD prevalence.

For the nonlinear PM2.5 effect, the posterior means are $\mathbb{E}[\beta_4] = -0.449$ and $\mathbb{E}[\beta_5] = 0.026$. This implies that, holding other predictors fixed, a ΔX_4^i-percentage-point increase in state PM2.5 is associated with a change in COPD prevalence of $(2\mathbb{E}[\beta_5]X_4^i + \mathbb{E}[\beta_4])\Delta X_4^i = (0.052X_4^i - 0.449)\Delta X_4^i$.

Table 3 presents posterior summaries for the hierarchical model under more diffuse priors on the standard deviations $\sigma_u, \sigma_Y \sim$ Half-Normal($\sigma = 5$). Priors of this scale are already quite broad relative to the empirical range of COPD prevalence, and therefore allow for larger variability in the state-specific effects and error terms. Even under this more permissive prior, the posterior estimates of all parameters change by no more than ± 0.1, providing strong evidence that the results from the original model are robust to prior specification.

Table 2: Summary of Posterior Estimates for Hierarchical Model Parameters across 48 states and DC (excluding Alaska and Hawaii) $[\sigma_u, \sigma_Y \sim \text{Half-Normal}(\sigma = 1)]$

Parameter	**Mean**	**SD**	**HDI 3%**	**HDI 97%**	**MCSE Mean**	**MCSE SD**	**ESS Bulk**	**ESS Tail**	$\hat{R}$
β_0	0.265	0.846	-1.347	1.812	0.027	0.015	953.0	1538.0	1.00
β_1	0.165	0.045	0.083	0.253	0.002	0.001	483.0	728.0	1.01
β_2	0.460	0.054	0.359	0.559	0.002	0.001	759.0	1863.0	1.00
β_3	0.066	0.047	-0.025	0.150	0.002	0.001	667.0	1389.0	1.01
β_4	-0.449	0.251	-0.934	0.012	0.010	0.006	663.0	1094.0	1.00
β_5	0.026	0.017	-0.005	0.059	0.001	0.000	687.0	976.0	1.00
σ_u	0.963	0.129	0.729	1.201	0.004	0.002	860.0	1791.0	1.01
σ_Y	0.513	0.032	0.455	0.576	0.001	0.000	1188.0	1655.0	1.01
State-specific Effects (u[State])									
u[AL]	0.724	0.355	0.046	1.395	0.015	0.007	596.0	1459.0	1.01
u[AR]	0.689	0.351	0.007	1.309	0.013	0.006	692.0	1498.0	1.01
u[AZ]	-0.004	0.318	-0.616	0.590	0.009	0.005	1375.0	2058.0	1.00
u[CA]	-1.200	0.436	-2.034	-0.387	0.016	0.008	780.0	1653.0	1.01
u[CO]	-0.245	0.343	-0.864	0.403	0.011	0.006	954.0	1749.0	1.00
u[CT]	-0.051	0.314	-0.647	0.544	0.010	0.005	976.0	1888.0	1.01
u[DC]	-0.399	0.404	-1.100	0.405	0.014	0.008	818.0	1640.0	1.01
u[DE]	0.243	0.291	-0.312	0.777	0.009	0.004	1055.0	2228.0	1.01
u[FL]	0.566	0.292	-0.012	1.091	0.008	0.005	1364.0	2221.0	1.01
u[GA]	-0.231	0.316	-0.814	0.374	0.009	0.005	1109.0	2047.0	1.00
u[IA]	-0.363	0.301	-0.920	0.214	0.009	0.005	1051.0	1994.0	1.01
u[ID]	-0.650	0.298	-1.197	-0.079	0.009	0.005	1233.0	1957.0	1.01
u[IL]	-0.221	0.306	-0.785	0.351	0.009	0.005	1272.0	2131.0	1.01
u[IN]	0.406	0.340	-0.249	1.031	0.014	0.006	608.0	1405.0	1.01
u[KS]	-0.404	0.290	-0.923	0.162	0.008	0.004	1288.0	2274.0	1.01
u[KY]	2.005	0.396	1.280	2.774	0.017	0.008	538.0	1361.0	1.01
u[LA]	-0.269	0.353	-0.972	0.348	0.012	0.005	890.0	1948.0	1.00
u[MA]	-0.125	0.315	-0.745	0.446	0.009	0.005	1128.0	1884.0	1.01
u[MD]	-0.558	0.320	-1.154	0.048	0.011	0.005	909.0	1920.0	1.01
u[ME]	0.919	0.328	0.327	1.543	0.010	0.005	1001.0	2011.0	1.00
u[MI]	0.987	0.302	0.400	1.535	0.010	0.005	971.0	1952.0	1.00
u[MN]	-1.010	0.316	-1.635	-0.441	0.010	0.005	988.0	1790.0	1.01
u[MO]	1.121	0.330	0.542	1.754	0.012	0.005	803.0	2142.0	1.01
u[MS]	-0.749	0.372	-1.410	-0.038	0.013	0.007	818.0	1433.0	1.00
u[MT]	0.569	0.311	0.013	1.188	0.008	0.005	1387.0	2178.0	1.00
u[NC]	0.133	0.286	-0.417	0.652	0.008	0.005	1381.0	2495.0	1.01
u[ND]	-1.287	0.339	-1.891	-0.620	0.010	0.005	1044.0	2038.0	1.00
u[NE]	-0.272	0.296	-0.813	0.315	0.008	0.005	1407.0	2157.0	1.01
u[NH]	0.733	0.375	0.016	1.413	0.013	0.006	872.0	1707.0	1.01
u[NJ]	0.019	0.343	-0.607	0.681	0.010	0.005	1135.0	2054.0	1.00
u[NM]	-1.893	0.441	-2.694	-1.064	0.015	0.007	875.0	1736.0	1.00
u[NV]	0.306	0.297	-0.268	0.866	0.008	0.005	1393.0	2227.0	1.01
u[NY]	-0.755	0.335	-1.376	-0.117	0.011	0.006	940.0	1904.0	1.01
u[OH]	0.518	0.349	-0.142	1.163	0.015	0.007	574.0	1075.0	1.01
u[OK]	0.266	0.306	-0.298	0.839	0.010	0.005	978.0	1918.0	1.00
u[OR]	0.062	0.285	-0.455	0.624	0.008	0.004	1393.0	2277.0	1.00
u[PA]	-0.015	0.295	-0.559	0.546	0.010	0.005	927.0	1843.0	1.01
u[RI]	0.797	0.294	0.284	1.383	0.007	0.005	1729.0	2270.0	1.00
u[SC]	-0.427	0.307	-0.987	0.165	0.010	0.005	954.0	1792.0	1.01
u[SD]	-1.775	0.323	-2.366	-1.132	0.009	0.005	1333.0	2239.0	1.00
u[TN]	0.934	0.336	0.346	1.623	0.014	0.006	572.0	1312.0	1.01
u[TX]	-1.826	0.335	-2.454	-1.201	0.010	0.006	1044.0	2007.0	1.00
u[UT]	0.034	0.402	-0.759	0.751	0.016	0.008	633.0	1174.0	1.01
u[VA]	0.054	0.298	-0.512	0.606	0.009	0.005	1196.0	2122.0	1.01
u[VT]	0.438	0.303	-0.105	1.037	0.008	0.005	1441.0	2208.0	1.00
u[WA]	-0.102	0.304	-0.676	0.476	0.009	0.005	1175.0	1967.0	1.01
u[WI]	-0.444	0.304	-0.986	0.153	0.009	0.006	1202.0	1960.0	1.01
u[WV]	3.150	0.464	2.284	4.023	0.021	0.010	476.0	1094.0	1.01
u[WY]	0.205	0.366	-0.493	0.878	0.011	0.006	1061.0	1961.0	1.00

Table 3: Summary of Posterior Estimates for Hierarchical Model Parameters across 48 states and DC (excluding Alaska and Hawaii) $[\sigma_u, \sigma_Y \sim \text{Half-Normal}(\sigma = 5)]$

Parameter	**Mean**	**SD**	**HDI 3%**	**HDI 97%**	**MCSE Mean**	**MCSE SD**	**ESS Bulk**	**ESS Tail**	$\hat{R}$
β_0	0.339	0.845	-1.160	1.948	0.029	0.015	878.0	1681.0	1.00
β_1	0.164	0.046	0.080	0.250	0.002	0.001	515.0	988.0	1.00
β_2	0.460	0.054	0.353	0.560	0.002	0.001	767.0	1589.0	1.00
β_3	0.068	0.045	-0.015	0.153	0.002	0.001	732.0	1502.0	1.00
β_4	-0.466	0.248	-0.952	-0.007	0.009	0.006	799.0	1191.0	1.00
β_5	0.027	0.017	-0.004	0.059	0.001	0.000	876.0	1278.0	1.00
σ_u	0.974	0.138	0.726	1.236	0.005	0.003	944.0	1372.0	1.00
σ_Y	0.513	0.032	0.455	0.575	0.001	0.000	1287.0	2482.0	1.00
State-specific Effects (u[State])									
u[AL]	0.711	0.357	0.042	1.386	0.012	0.008	827.0	1169.0	1.00
u[AR]	0.676	0.357	-0.041	1.319	0.012	0.007	850.0	1075.0	1.00
u[AZ]	-0.016	0.313	-0.618	0.560	0.010	0.005	969.0	2108.0	1.00
u[CA]	-1.235	0.438	-2.045	-0.413	0.016	0.008	758.0	1504.0	1.00
u[CO]	-0.251	0.339	-0.898	0.336	0.013	0.007	696.0	1381.0	1.01
u[CT]	-0.052	0.327	-0.661	0.572	0.011	0.007	830.0	1284.0	1.00
u[DC]	-0.415	0.407	-1.169	0.355	0.015	0.008	781.0	1380.0	1.00
u[DE]	0.242	0.294	-0.312	0.785	0.009	0.005	1093.0	2023.0	1.00
u[FL]	0.558	0.290	0.038	1.124	0.008	0.005	1176.0	1988.0	1.00
u[GA]	-0.246	0.312	-0.875	0.299	0.009	0.005	1155.0	1759.0	1.00
u[IA]	-0.367	0.305	-0.926	0.220	0.010	0.005	920.0	1715.0	1.00
u[ID]	-0.665	0.307	-1.251	-0.098	0.011	0.006	845.0	1572.0	1.00
u[IL]	-0.231	0.316	-0.820	0.360	0.010	0.005	1050.0	1748.0	1.00
u[IN]	0.402	0.349	-0.229	1.075	0.012	0.007	911.0	1175.0	1.00
u[KS]	-0.414	0.292	-0.937	0.154	0.009	0.004	1101.0	2210.0	1.00
u[KY]	1.998	0.404	1.258	2.774	0.015	0.009	726.0	971.0	1.00
u[LA]	-0.282	0.356	-0.912	0.410	0.012	0.006	948.0	1437.0	1.00
u[MA]	-0.135	0.329	-0.783	0.444	0.012	0.006	790.0	1265.0	1.01
u[MD]	-0.561	0.323	-1.156	0.042	0.011	0.005	827.0	1303.0	1.00
u[ME]	0.917	0.327	0.276	1.517	0.010	0.007	985.0	1271.0	1.00
u[MI]	0.986	0.305	0.405	1.546	0.009	0.005	1209.0	1869.0	1.00
u[MN]	-1.017	0.319	-1.603	-0.418	0.012	0.006	729.0	1208.0	1.01
u[MO]	1.113	0.332	0.523	1.761	0.012	0.006	826.0	1440.0	1.00
u[MS]	-0.770	0.375	-1.467	-0.056	0.013	0.007	828.0	1370.0	1.00
u[MT]	0.564	0.311	-0.048	1.134	0.010	0.005	1055.0	2080.0	1.00
u[NC]	0.123	0.281	-0.406	0.665	0.008	0.004	1251.0	1829.0	1.00
u[ND]	-1.290	0.336	-1.931	-0.665	0.011	0.007	874.0	1313.0	1.00
u[NE]	-0.274	0.300	-0.832	0.283	0.010	0.005	971.0	1617.0	1.00
u[NH]	0.734	0.368	0.023	1.391	0.013	0.007	838.0	1349.0	1.00
u[NJ]	0.011	0.354	-0.688	0.640	0.011	0.006	1012.0	1660.0	1.00
u[NM]	-1.918	0.440	-2.724	-1.094	0.014	0.007	929.0	2083.0	1.00
u[NV]	0.297	0.292	-0.254	0.851	0.008	0.005	1337.0	2149.0	1.00
u[NY]	-0.765	0.333	-1.400	-0.129	0.012	0.006	818.0	1603.0	1.00
u[OH]	0.511	0.351	-0.138	1.172	0.012	0.007	890.0	1544.0	1.00
u[OK]	0.247	0.308	-0.300	0.838	0.009	0.005	1123.0	1850.0	1.00
u[OR]	0.052	0.293	-0.483	0.603	0.009	0.005	1013.0	1410.0	1.00
u[PA]	-0.015	0.306	-0.580	0.584	0.009	0.005	1144.0	2062.0	1.00
u[RI]	0.789	0.298	0.234	1.341	0.009	0.005	1122.0	2031.0	1.00
u[SC]	-0.444	0.309	-1.060	0.080	0.009	0.006	1078.0	1739.0	1.00
u[SD]	-1.775	0.319	-2.351	-1.169	0.009	0.006	1161.0	1734.0	1.00
u[TN]	0.925	0.345	0.273	1.558	0.012	0.007	870.0	1192.0	1.00
u[TX]	-1.843	0.332	-2.491	-1.225	0.011	0.006	907.0	1640.0	1.00
u[UT]	0.020	0.427	-0.737	0.849	0.017	0.010	621.0	829.0	1.01
u[VA]	0.048	0.300	-0.480	0.627	0.009	0.005	1007.0	2068.0	1.00
u[VT]	0.425	0.307	-0.163	0.993	0.010	0.005	998.0	1930.0	1.00
u[WA]	-0.112	0.317	-0.697	0.483	0.011	0.006	876.0	1406.0	1.00
u[WI]	-0.450	0.315	-1.039	0.134	0.011	0.006	853.0	1190.0	1.00
u[WV]	3.148	0.479	2.229	4.029	0.019	0.012	651.0	961.0	1.00
u[WY]	0.194	0.367	-0.505	0.857	0.012	0.007	1009.0	1453.0	1.00

5.0.3 Discussion

As illustrated in Figure 6, the posterior estimates of the state-specific effects exhibit notable heterogeneity, ranging from negative to positive values. Although the hierarchical model does not identify which unobserved covariates drive these differences, the patterns are consistent with factors documented in prior epidemiological and environmental research.

One plausible contributor is the proportion of Hispanic residents in each state. Prior work shows that, among individuals who smoke, Hispanic adults exhibit lower rates of airflow obstruction relative to non-Hispanic White adults [16]. States such as Texas and New Mexico, which have some of the highest Hispanic population shares in the United States [17], show large negative state-specific effects, around a 2-percentage-point reduction in COPD prevalence relative to model predictions.

The other plausible covariates are environmental legacies of historical coal extraction [18] and the proportion of coal miners who were employed before the implementation of the Federal Mine Safety and Health Act of 1977 [19]. Before this legislation mandated comprehensive safety standards and regular inspections, miners were exposed to substantially higher levels of occupational hazards, including respirable dust associated with chronic lung disease. States such as West Virginia and Kentucky, located in Appalachia, show state-level effects that elevate predicted COPD prevalence by approximately 2%. This pattern is consistent with the region's long-standing history of intensive coal mining. From the 1800s through the 1970s, Appalachia served as the nation's primary coal-producing region, and the cumulative environmental and occupational impacts of this century-scale industry have shaped the area's landscapes, ecosystems, and public health outcomes for generations [20].

Incorporating data on state-level ethnic composition and the prevalence of coal miners could therefore improve model robustness.

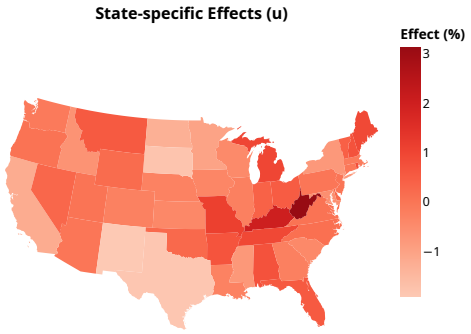


Figure 6: Posterior mean estimates of state-specific effects u_i

Our analysis confirms a positive association between state-level smoking rates and COPD prevalence, consistent with previous findings [7]. While prior research implies a linear association between PM2.5 concentration and COPD prevalence in Taiwan [21], the results presented in Figure 4 and Table 2 suggest a potential quadratic relationship between PM2.5 levels and COPD prevalence in the United States.

Nevertheless, this apparent nonlinearity does not necessarily imply a true quadratic effect. The adverse health impacts of PM2.5 are generally observable only above a certain threshold [22]. Indeed, the subset of data in Figure 4 where $X_4^i \lesssim 8.635$ exhibits no meaningful correlation between PM2.5 concentration and COPD prevalence. Consequently, the observed quadratic pattern is likely driven by confounding factors that influence both PM2.5 levels and COPD prevalence, rather than a direct nonlinear effect of PM2.5 itself.

Finally, it is important to note that Bayesian hierarchical models estimate associations rather than causal effects, so this study cannot make definitive claims about causality.

6 Predicting 2020 COPD Prevalence: Bayesian Hierarchical Model vs. Naive Linear Regression

6.0.1 Methods

`COPD_rate` contains the actual observed 2020 state-level COPD prevalence data. In this section, we evaluate the predictive performance of a Bayesian Hierarchical Model and a Naive Linear Regression by forecasting the 2020 COPD prevalence and calculating the root mean squared error (RMSE) as a measure of model accuracy.

For the Bayesian Hierarchical Model, we first generate 2020 state-level predictions for smoking prevalence, diabetes prevalence, poverty rate, and $PM_{2.5}$ concentration by applying Algorithm 1 to the respective datasets (`Smoking_rate`, `Diabetes_rate`, `Poverty_rate`, and `PM2.5_level`). Using these predicted covariates, the 2020 state-level COPD prevalence Y_i is then estimated as

$$Y_i = \mathbb{E}[u_i] + \mathbb{E}[\beta_0] + \mathbb{E}[\beta_1]X_1^i + \mathbb{E}[\beta_2]X_2^i + \mathbb{E}[\beta_3]X_3^i + \mathbb{E}[\beta_4]X_4^i + \mathbb{E}[\beta_5](X_4^i)^2$$

For the Naive Linear Regression, it predicts 2020 COPD prevalence directly by applying Algorithm 1 to the `COPD_rate` dataset, without incorporating additional covariates or hierarchical structure.

Algorithm 1 Extrapolate Values to 2020

Require: data grouped by states
Ensure: predicted values for 2020
Initialize empty list `results`
for each state in data **do**
  Get list of years and values
  if number of years > 2 **then**
    Fit a line to (years, values)
    Predict value for 2020 using the line
  else
    Predict value for 2020 as the average of values
  end if
  Append state and predicted 2020 value to `results`
end for
return `results`

6.0.2 Results and Discussion

The Root Mean Square Error (RMSE) for the Bayesian Hierarchical Model was 0.5617, compared to 0.7767 for the Naive Linear Regression. The Bayesian Hierarchical Model more effectively incorporates latent variations associated with multiple relevant covariates that the Naive Linear Regression fails to capture, thereby enhancing predictive accuracy. This result shows the predictive advantage of integrating information from multiple sources.

7 Conclusion

This study analyzed adult COPD prevalence at the U.S. state level and its relationship with smoking rates, diabetes rates, poverty rates, and PM2.5 concentrations using a Bayesian hierarchical model. The analysis found that smoking and diabetes rates are positively correlated with COPD prevalence, with a 94% credible interval. Poverty rates were also found to have a positive correlation with COPD prevalence, suggesting that socioeconomic disparities might contribute to respiratory disease risk. Additionally, PM2.5 concentrations suggest a possibility of a non-linear association between COPD prevalence, with low concentrations having minimal impact, while high concentrations significantly increase COPD prevalence. State-level effects may reflect factors such as historical industry patterns and ethnic composition, showing the importance of accounting for region-specific health risks.

There are several limitations in this analysis. The BRFSS survey data are based on telephone interviews and may underrepresent individuals without phone access. Self-reported information on smoking and diabetes may also be affected by recall or social desirability bias. Furthermore, PM2.5 data are aggregated at the state level, which may not fully capture individual-level exposure differences. Some domain knowledge, like environmental epidemiology, was limited and constrained variable selection and model specification. For example, asking a domain expert how occupational exposures influence state-level COPD prevalence could have informed the choice of covariates and improved model accuracy. Regarding robustness, varying the priors of σ_u, σ_Y in the hierarchical model did not substantially change the parameter estimates, indicating stability. However, assuming a linear effect of PM2.5 may underestimate its impact in regions with high concentrations. Since the study uses state-level aggregated data, the findings cannot directly infer individual-level causation, but they are broadly applicable for state-level public health policy and regional interventions.

Based on these findings, future studies could incorporate ethnic distribution and occupation-specific risk factors in hierarchical models. Including historical industry or mining legacy variables could also clarify state-specific latent effects. In terms of policy recommendations, state governments and environmental agencies should prioritize improving air quality in regions with high PM2.5 levels (especially the region with an average yearly concentration above $8.635\mu g/m^3$). States with high smoking or diabetes rates should implement stronger health education, smoking cessation programs, and lifestyle interventions. The feasibility of these recommendations largely depends on government and agency authority, although industrial stakeholders may object. The impacts of such intervention on air quality would vary between individuals and groups, with high-risk residents benefiting from improved health outcomes, while some individuals, like industrial workers, might experience economic or occupational negative effects. Despite these potential trade-offs, these recommendations need serious consideration, since they are guided by the ethical values of promoting human health and evidence from the data.

References

[1] Yildirim, K. 2023. U.S. Chronic Disease Indicators (CDI) — 2023 Release [Dataset]. Kaggle. `https://www.kaggle.com/datasets/kadiryildirim12/u-s-chronic-disease-indicators-cdi-2023-release/data`. Accessed 1 Dec. 2025.

[2] Centers for Disease Control and Prevention. n.d. Behavioral Risk Factor Surveillance System (BRFSS) Prevalence Data (2011 to present) [Dataset]. `https://data.cdc.gov/Behavioral-Risk-Factors/Behavioral-Risk-Factor-Surveillance-System-BRFSS-P/dttw-5yxu/about_data`. Accessed 1 Dec. 2025.

[3] Kaiser Family Foundation. 2023. Poverty Rate by Race/Ethnicity [Dataset]. `https://www.kff.org/state-health-policy-data/state-indicator/poverty-rate-by-raceethnicity/`. Accessed 1 Dec. 2025.

[4] Centers for Disease Control and Prevention. n.d. Daily Census Tract-Level $PM_{2.5}$ Concentrations, 2016–2020 [Dataset]. https://data.cdc.gov/Environmental-Health-Toxicology/Daily-Census-Tract-Level-PM2-5-Concentrations-2016/96sd-hxdt/about_data. Accessed 1 Dec. 2025.

[5] U.S. Census Bureau. 2010. Chapter 11. Weighting and Estimation [Revised December 2010]. https://www.census.gov/content/dam/Census/library/publications/2010/acs/Chapter_11_RevisedDec2010.pdf. Accessed 14 Dec. 2025.

[6] Centers for Disease Control and Prevention. 2020. Weighting the BRFSS Data. https://www.cdc.gov/brfss/annual_data/2020/pdf/weighting-2020-508.pdf. Accessed 14 Dec. 2025.

[7] Terzikhan, N., Verhamme, K. M. C., Hofman, A., Stricker, B. H., Brusselle, G. G., and Lahousse, L. 2016. "Prevalence and incidence of COPD in smokers and non-smokers: the Rotterdam Study." *European Journal of Epidemiology* **31**(8): 831–842. https://pmc.ncbi.nlm.nih.gov/articles/PMC5005388/. Accessed 1 Dec. 2025.

[8] Chung, C., Lee, K. N., Han, K., Shin, D. W., and Lee, S. W. 2023. "Effect of smoking on the development of chronic obstructive pulmonary disease in young individuals: a nationwide cohort study." *Frontiers in Medicine* **10**. https://pmc.ncbi.nlm.nih.gov/articles/PMC10428618/. Accessed 1 Dec. 2025.

[9] Hikichi, M., Mizumura, K., Maruoka, S., and Gon, Y. 2019. "Pathogenesis of chronic obstructive pulmonary disease (COPD) induced by cigarette smoke." *Journal of Thoracic Disease* **11**(Suppl 17): S2129–S2135. https://pmc.ncbi.nlm.nih.gov/articles/PMC6831915/. Accessed 1 Dec. 2025.

[10] Gong, E., Kou, Z., Li, Y., Li, Q., Yu, X., Wang, T., and Han, W. 2025. "Weight change across adulthood in relation to the risk of COPD." *Environmental Health and Preventive Medicine* **30**: 64. https://pmc.ncbi.nlm.nih.gov/articles/PMC12358758/. Accessed 1 Dec. 2025.

[11] Zhang, S.-J., Qin, X.-Z., Zhou, J., He, B.-F., Shrestha, S., Zhang, J., and Hu, W.-P. 2023. "Adipocyte dysfunction promotes lung inflammation and aberrant repair: a potential target of COPD." *Frontiers in Endocrinology* **14**: 1204744. https://www.frontiersin.org/journals/endocrinology/articles/10.3389/fendo.2023.1204744/. Accessed 1 Dec. 2025.

[12] Lee, Y. S., Oh, J. Y., Min, K. H., Lee, S. Y., Kang, K. H., and Shim, J. J. 2019. "The association between living below the relative poverty line and the prevalence of chronic obstructive pulmonary disease." *Journal of Thoracic Disease* **11**(2): 399–407. https://pmc.ncbi.nlm.nih.gov/articles/PMC6409249. Accessed 1 Dec. 2025.

[13] Vestbo, J., Hurd, S. S., Agustí, A. G., Jones, P. W., Vogelmeier, C., Anzueto, A., Barnes, P. J., Fabbri, L. M., Martinez, F. J., Nishimura, M., Stockley, R. A., Sin, D. D., and Rodriguez-Roisin, R. 2013. "Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Pulmonary Disease: GOLD Executive Summary." *American Journal of Respiratory and Critical Care Medicine* **187**(4): 347–365. https://www.atsjournals.org/doi/10.1164/rccm.201204-0596PP. Accessed 1 Dec. 2025.

[14] Bo, Y., Chang, L.-Y., Guo, C., Lin, C., Lau, A. K. H., Tam, T., and Lao, X. Q. 2021. "Reduced ambient $PM_{2.5}$, better lung function, and decreased risk of chronic obstructive pulmonary disease." *Environment International* **156**: 106706. https://www.sciencedirect.com/science/article/pii/S0160412021003317. Accessed 1 Dec. 2025.

[15] Li, T., Yu, Y., Sun, Z., and Duan, J. 2022. "A comprehensive understanding of ambient particulate matter and its components on the adverse health effects based from epidemiological and laboratory evidence." *Particle and Fibre Toxicology* **19**: 62. https://pmc.ncbi.nlm.nih.gov/articles/PMC9707232/. Accessed 1 Dec. 2025.

[16] Flores, G., et al. 2022. "Ethnic Differences in Airflow Obstruction Among U.S. Adults Who Smoke: NHANES 2007–2012." *COPD: Journal of Chronic Obstructive Pulmonary Disease* **19**(1): 1–10. https://www.tandfonline.com/doi/full/10.1080/15412555.2022.2029384. Accessed 11 Dec. 2025.

[17] Pew Research Center. 2013. "Mapping the Latino Population, By State, County and City." https://www.pewresearch.org/race-and-ethnicity/2013/08/29/mapping-the-latino-population-by-state-county-and-city/. Accessed 11 Dec. 2025.

[18] Gopinathan, P., Subramani, T., Barbosa, S., and Yuvaraj, D. 2023. "Environmental impact and health risk assessment due to coal mining and utilization." *Environmental Geochemistry and Health* **45**(12): 8855–8875. https://link.springer.com/article/10.1007/s10653-023-01744-z. Accessed 11 Dec. 2025.

[19] Mine Safety and Health Administration. 2016. *Mine Safety and Health Deskbook*. https://www.msha.gov/sites/default/files/Regulations/mine-safety-and-health-deskbook.pdf. Accessed 11 Dec. 2025.

[20] Zipper, C. E., Adams, M. B., and Skousen, J. 2021. "The Appalachian coalfield in historical context." In *Appalachian set aside and its legacy*, edited by C. E. Zipper and P. F. Ziemkiewicz, 1–10. Gen. Tech. Rep. NRS-P-195. Madison, WI: U.S. Department of Agriculture, Forest Service, Northern Research Station. https://research.fs.usda.gov/treesearch/62623. Accessed 11 Dec. 2025.

[21] Bo, Y., Chang, L.-Y., Guo, C., Lin, C., Lau, A. K. H., Tam, T., and Lao, X. Q. 2021. "Reduced ambient $PM_{2.5}$, better lung function, and decreased risk of chronic obstructive pulmonary disease." *Environment International* **156**: 106706. https://www.sciencedirect.com/science/article/pii/S0160412021003317. Accessed 11 Dec. 2025.

[22] Guo, Q., Zhao, Y., Zhao, J., Qian, L., Bian, M., Xue, T., Zhang, J., and Duan, X. 2022. "Identifying the threshold of outdoor $PM_{2.5}$ reversing the beneficial association between physical activity and lung function: A national longitudinal study in China." *Science of The Total Environment* **839**: 155938. https://www.sciencedirect.com/science/article/abs/pii/S0048969722032351. Accessed 11 Dec. 2025.