Breathing the Divide: Modeling COPD Prevalence, Air Pollution, and Smoking Across U.S. States

Albert Ponferrada Sicart

Data Overview

The primary dataset that we used was the Center for Disease Control and Prevention's (CDC) *U.S. Chronic Disease Indicators* dataset, which contains state and territory-specific data of chronic illness prevalence. This dataset compiles data from fourteen different sources such as sample surveys, census data and registry data. We accessed this data using the API to help us with managing the large dataset; we set a topic parameter for 'Chronic Obstructive Pulmonary Disease' (COPD) to filter directly to suit our research questions. Initial pre-processing to our dataframe was making the year and data value features numeric. Our interest was in per-state values, so each row in our dataset represents each state; we believe that this granularity is sufficient for general results, however a finer granularity such as county-level may have been more useful when examining other demographic information. We chose to use the prevalence data type, and when filtering for this the data source was 'BRFSS' or the Behavioral Risk Factor Surveillance System. These are health-related telephone surveys, collecting data from US residents about health-related risk behaviors, chronic health conditions, health-care access and use of preventive services (CDC, 2024). As a result, the dataset cannot be treated as a census. Within the data value type of our dataset, there is a 'Crude prevalence' and a 'Age-adjusted prevalence' type, which allows us to compare how these values are different. When visualizing these two types per state, almost every state had a lower age-adjusted prevalence with only three states having a higher value. Using the age-adjusted value instead of crude prevalence could help with the generalizability of our results, because we are standardizing the age distribution throughout the states. This could also help with some concerns in our data such as selection bias; because the method of collecting data is through phone calls, this could lead to a specific age group that respond or provide contact information, then adjusting for age could remove any specific weighting for an age group for each state. After applying the data value type and topic, we did not have any of our meaningful features with missing values.

Our analysis would be enhanced by several variables but this was not available. Personal BRFSS microdata would allow making the analyses that would maintain control over personal smoking history, occupation, and residential length and minimize ecological fallacy concerns. As discussed earlier, relying on 2020-only data limits our ability to really tie exposure with COPD's long-term development.

The measures of access to healthcare such as coverage rates of insurance and pulmonologist density per capita and availability of spirometry testing measurements would be

used to differentiate between the actual COPD burden and the rates of the differential diagnosis in the states. Non-ambient respiratory hazards would be managed by occupational exposure information of state level occupational employment in higher risk trades such as mining, manufacturing, and agriculture. The measurements of indoor air quality are also not present, and since PM2.5 measurements indicate the outdoor ambient air, but do not account for indoor exposures (cooking, heating, also secondhand smoke), these figures might be significant especially in low-income groups.

Another main dataset we used was the CDC's *Daily Census Tract-Level PM2.5 Concentrations* dataset, from 2016-2020. This dataset has predictions of PM2.5 (fine particulate matter in the air) levels in $\mu g/m^3$ at the census tract level. This gave us county level predictions, providing census-like data. We then grouped by state to achieve a state-wide average to be usable with our COPD dataset. Doing this on the CDC website allowed us to download the dataset because we had reduced the number of rows and therefore size. Using this group method does raise a measurement error concern in terms of the standard error going from county level to state level, as well as making generalizations for whole states, when particular counties may cause a higher prediction.

Other demographic information was collected from the US Census Bureau and CDC website. We used the populations per state from the 2020 census, COPD and tobacco data from the API to get data about current smoking levels among adults and adults with COPD per state, median state income and household type data. These extra datasets were particularly useful in our EDA, visualizing different potential relationships.

Research Questions

With our first research question, we are interested in how COPD prevalence rates vary across different U.S. states and demographic subgroups, and if we can use Bayesian hierarchical modeling to understand national-level variation whilst accounting for regional differences in air quality. Exploring this could highlight states that need further policy changes or prioritization of improving air quality, using COPD cases as a strong justification. Similarly, results from this question could lead to a shift in resource allocation, and how states and the U.S. can focus on specific areas or demographics that show a high prevalence. Bayesian hierarchical modeling is a good fit for our question because there is a clear hierarchical structure; within the United States, we have regional air-qualities, within these we have state data which have demographic subgroups. This type of modeling allows us to use partial pooling for smaller states, account for and separate explained vs. unexplained differences, for example air quality regionally is likely to be similar. Partial pooling allows us to share this information across regions, preventing both less accurate results, and overfitting. Some of the limitations using this model is that we are assuming heterogeneity within the state, and so some of the state averages might accurately represent specific urban or rural areas within the state.

Our second research question is looking at the causal effect of PM2.5 concentrations. More specifically, does exposure to elevated PM2.5 increase COPD rates at the state level, after controlling for confounders such as demographics, socioeconomic and behavioral. Similar to our first research question, policy and regulation decisions could be made if there is a causal association between COPD and PM2.5 concentrations where stricter standards could be enforced. Another avenue could be to release more publicly distributed information about this causal relationship, spreading more education and information about what can contribute to COPD as well as how air quality is important. Causal inference is a suitable method for this question because we are looking to explore a causal relationship, not just an association between COPD and PM2.5. This method also allows us to account for confounding variables, and balance weights for different groups with propensity scores, however the main limitation is accounting for sufficient confounders; if we don't have enough data on some confounders as well as missing crucial variables out, our model is unlikely to be useful for causal effects.

Prior Work

In Liu et al.'s report for the CDC, the study examines temporal trends in COPD prevalence across U.S. states and demographic subgroups from 2011-2021 using BRFSS data (same as our dataset above). This question parallels our first research question about COPD prevalence varying across states and groups. While the study finds smoking as the dominant COPD risk factor, they find that 25% occur among people who have never smoked, providing rationale for our second research question, investigating air quality factor PM2.5. BRFSS data is used in both analyses, using age-adjusted prevalence, examining different demographic disparities, but the report uses data from 2011 to 2021 where we use 2020 data. They took a more frequentist approach using a jointplot regression analysis where we used a Bayesian hierarchical model. Their more sophisticated model could detect non-linear temporal trends where ours cannot. Our model however, is able to handle multi-level structures better.

Park et al.'s review and meta-analysis, they pooled high-quality evidence from previous studies to estimate the long term effects of air pollution on the incidence of COPD by PM and CO_2. This is related to our causal inference question, because we are also interested in the effects of particulate matter on COPD. Through these studies, they calculated the standardized effect estimates indicating risk of COPD development per 10 $\mu g/m^3$ increase in specific air pollutants, which were used to calculate pooled estimates using a random-effects model. Their units were individual participants through different regions compared to our unit of states in the U.S. Their results found that out of PM2.5, NO_2 and PM10, PM2.5 exposure had the strongest association with the incidence of COPD. A 10 $\mu g/m^3$ increase in PM2.5 is associated with an 18% increased incidence or hazard ratio of 1.18 for COPD (Park et al., 2021). Although their method of meta-analysis and meta-regression is not very similar to ours, and that they were exploring association between variables, this study acts as a strong motivator to our research question, to see if there is not just an association but a causation. This study had access to more data, and so it is likely that their results could be more specific or accurate, especially because they were

analyzing at an individual level. One part to note is their limitation that COPD diagnoses were different depending on the study they were inspecting; this is different to our data collection, where every prevalence rate is collected from the BRFSS.

EDA

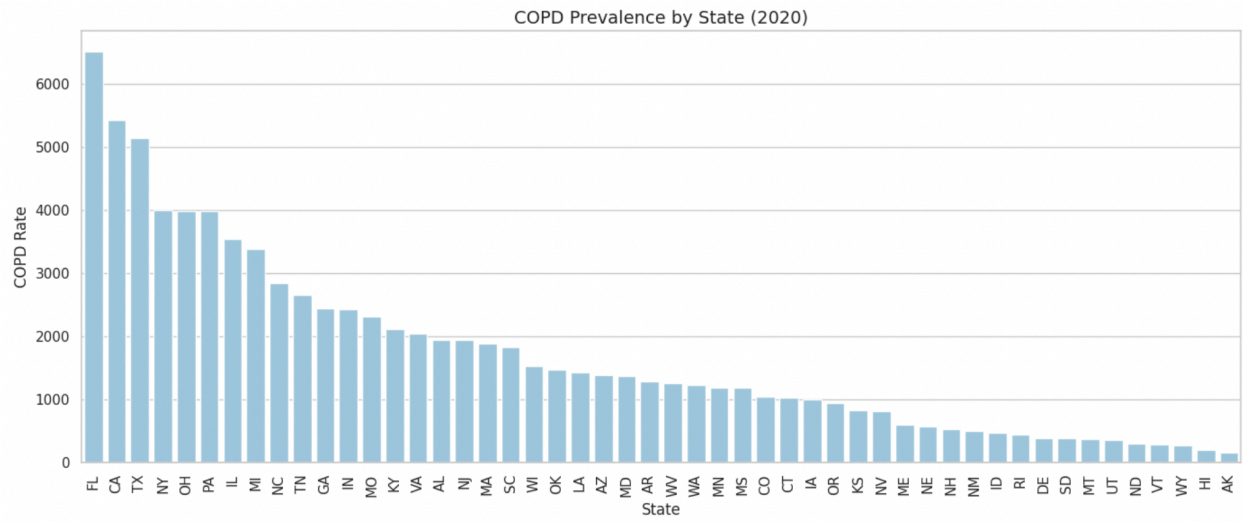
To better understand patterns in COPD prevalence and inform our hierarchical modeling approach, we conducted exploratory data analysis using multiple datasets, including COPD prevalence, PM2.5 air pollution, smoking behavior, and demographic stratifications. All analyses were restricted to **U.S. states in the year 2020** to ensure temporal and geographic consistency across datasets.

For the COPD dataset, we explored prevalence across demographic subgroups using the 'stratificationcategory1' variable. Separate subsets were created for gender and race/ethnicity. The gender subset included only observations labeled "Male" and "Female." Race and ethnicity categories were standardized by mapping long or inconsistent labels (e.g., "Hawaiian or Pacific Islander, non-Hispanic") to concise, uniform labels (e.g., "HI/PI") to improve interpretability in visualizations.

Missing or invalid prevalence values were removed prior to aggregation. For national-level summaries, COPD prevalence was grouped by stratification category. For state-level analyses, data were grouped by state and stratification to compute average prevalence rates.

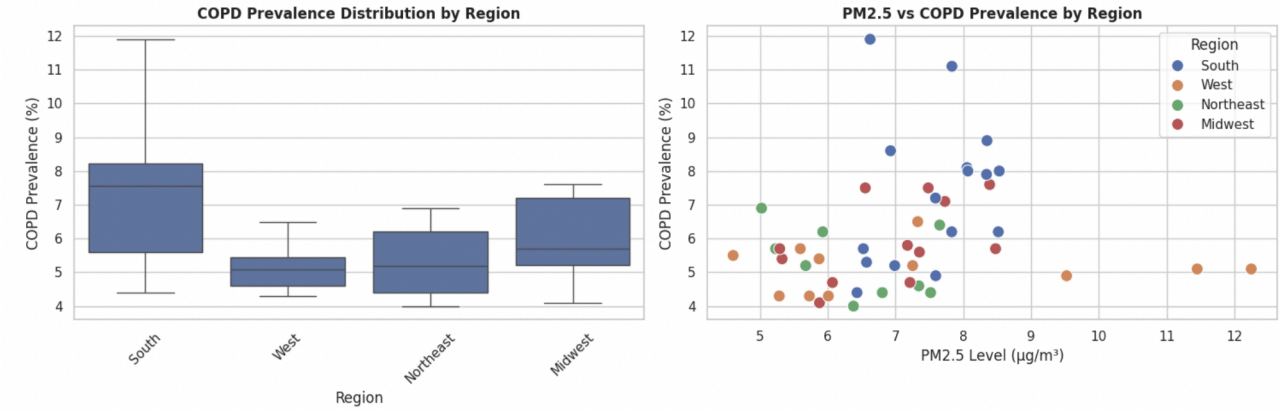
For the PM2.5 dataset, daily predicted concentrations were aggregated to obtain state-level average PM2.5 exposure for 2020. State FIPS codes were mapped to state abbreviations to allow merging with the COPD dataset. Smoking prevalence data were constructed using two CDC indicators: (1) cigarette smoking among all adults and (2) cigarette smoking among adults with COPD. These indicators were aggregated to the state level, allowing comparison between general smoking behavior and smoking prevalence within the COPD population.

Figure 1. Bar chart of COPD prevalence by state in 2020



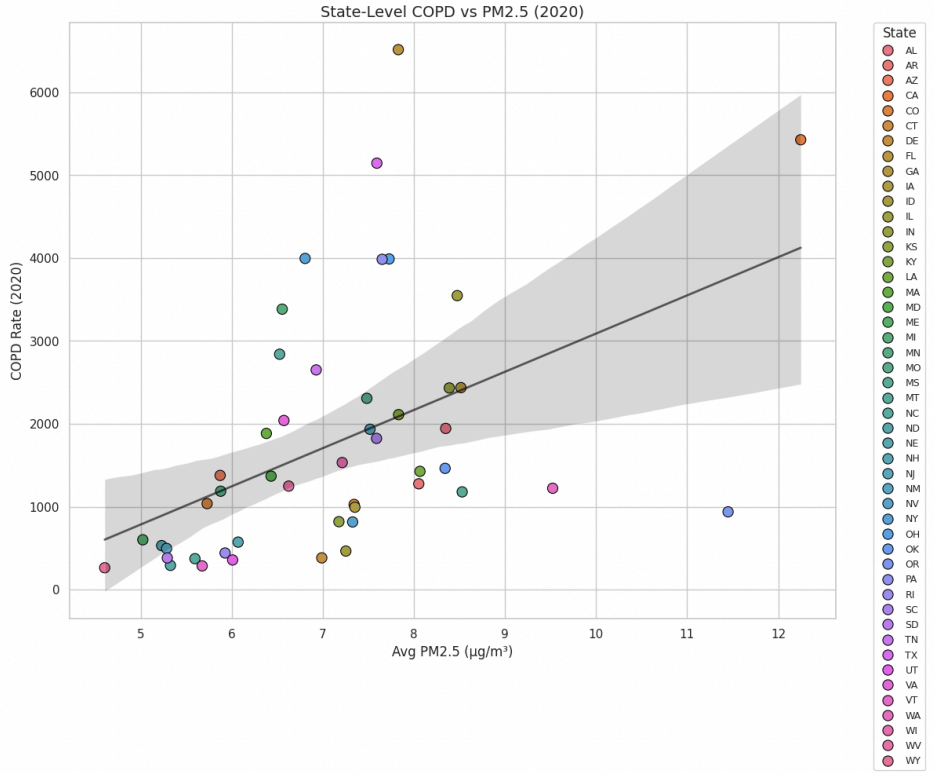
This visualization highlights substantial variation in COPD prevalence across states. States such as Florida, California, and Texas exhibit notably higher prevalence rates, which is consistent with patterns often observed in larger and more populous states.

Figures 2 & 3. Box plot of COPD prevalence by region (left), COPD prevalence vs. PM2.5 concentrations by region (right)



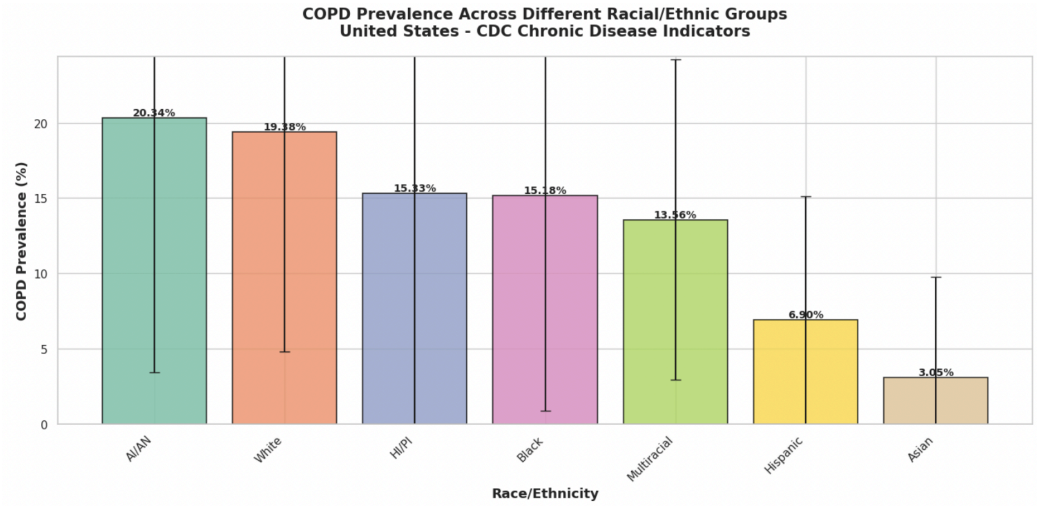
These plots show clear regional differences in COPD prevalence and PM2.5 exposure. Both COPD rates and PM2.5 concentrations vary substantially across regions, suggesting that geographic structure may play an important role in explaining observed prevalence patterns.

Figure 4. Scatter plot of COPD rate vs average PM2.5 concentration by state in 2020



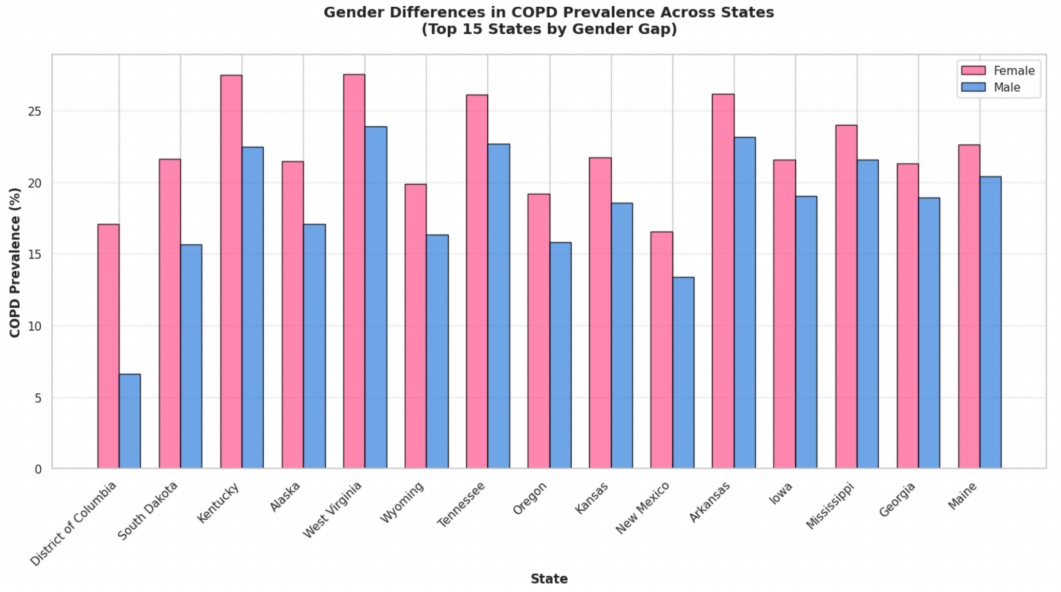
The scatter plot above directly relates to our causal inference research question - it shows a weak positive association between PM2.5 exposure and COPD prevalence at the state level. While states with higher PM2.5 concentrations tend to have higher COPD prevalence, the relationship is not strong and several outliers are present.

Figure 5. Boxplot of COPD prevalence across racial and ethnic groups



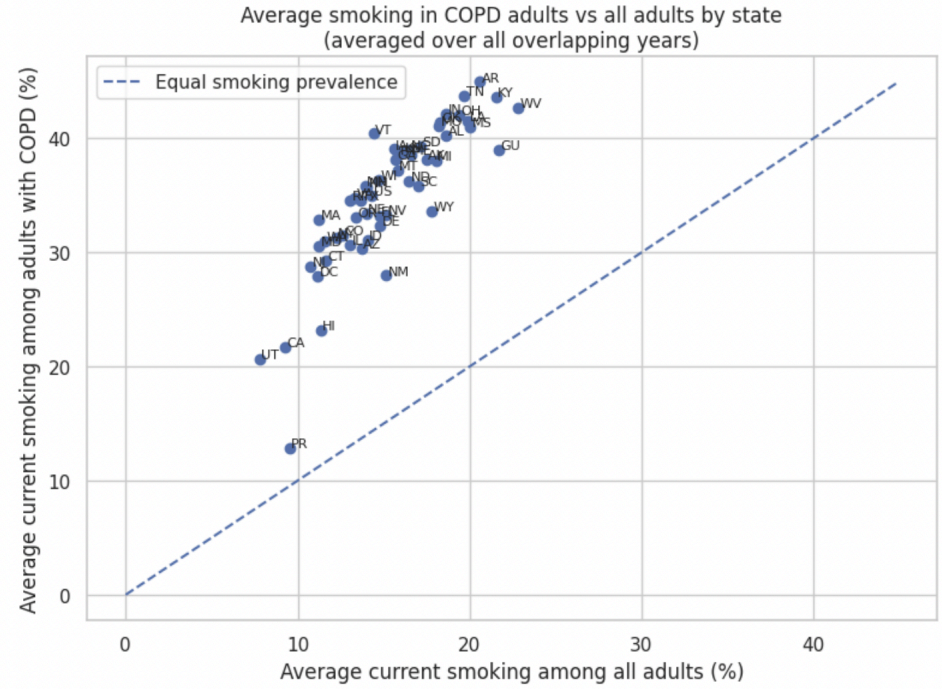
This boxplot illustrates that COPD prevalence is highest among Native American and White populations and lowest among Asian and Hispanic populations, with substantial variability across groups.

Figure 6. Grouped bar chart of female and male COPD prevalence in the top 15 states



Here, we can see that across nearly all states shown, female COPD prevalence exceeds male prevalence, and this pattern is relatively consistent.

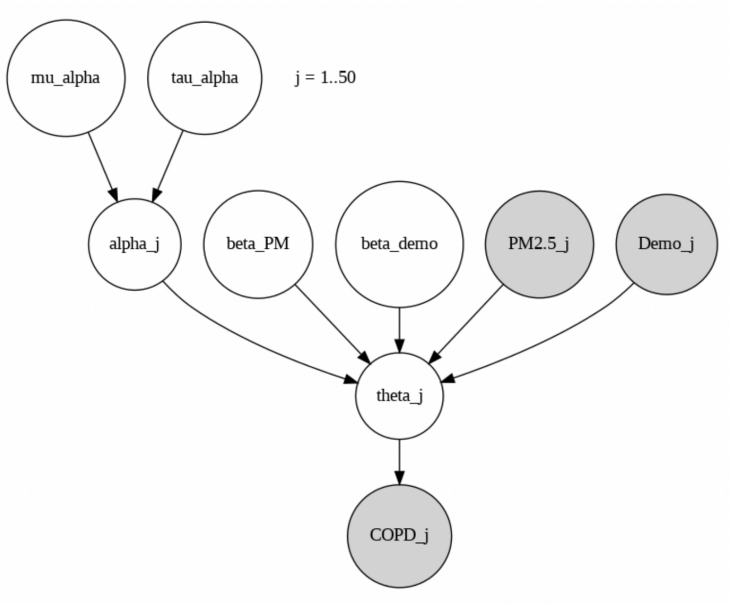
Figure 7. Average smoking in adults with COPD vs. average smoking among all adults by state



Here, nearly all states lie above the equality line, indicating that smoking prevalence is substantially higher among adults with COPD than in the general population.

Research Question 1

Methods



μ_a: the global mean COPD prevalence across all states
τ_a: the global variability in state-level baseline prevalence
α□: the latent baseline COPD prevalence for state j, capturing unobserved state-specific factors such as healthcare access, climate, and long-term behavioral patterns.
β_PM: the global effect of PM2.5 exposure on COPD prevalence.
β_demo: the effect of observed demographic covariates.
PM2.5□: avg. PM2.5 concentration for state j
Demo□: demographic characteristics for state j
θ□: the latent expected COPD prevalence for state j, modeled as a function of the state intercept and observed covariates.
COPD□: observed age-adjusted COPD prevalence for state j, generated from the latent prevalence θj.

The shaded nodes are observed variables; the unshaded nodes are latent parameters estimated by the model.

From the datasets previously mentioned, we used age-adjusted COPD prevalence, PM2.5 concentration predictions by state, gender and ethnicity measures, and the regions each state lies in. We used the year 2020 with 48 states of full information, with our predictors being standardized for easier interpretability. Our exploratory data analysis helped us make some decisions for our model. Firstly, including the regional random effects which accounts for

clustering of states within regions. Secondly, was that PM2.5 effects vary by region, which capture differing environmental impacts. Finally, we include gender gap as a fixed effect to test whether demographic disparities predict overall state prevalence.

We implemented a two-level Bayesian hierarchical model, with our first level being state-level observations and our second level being regional parameters. For each state i in region j, our observed age-adjusted COPD prevalence follows a normal distribution where expected prevalence is a linear function of regional COPD baseline for the region, allowing air quality change across regions, and for gender disparity. The first term captures all unmeasured factors that make states in the region similar (e.g. climate), the second term reflects heterogeneous environments and air quality regulations. The third term is seeing if states with larger gender COPD disparities have overall different levels.

```
mu = (alpha_region[region_idx] +  beta_pm25_region[region_idx] *
            pm25 +  beta_gender * gender_gap)
```

For our regional estimates, we chose to implement partial pooling instead of estimating each regional parameter to avoid overfitting. The regional estimates are ‘shrunk’ toward the population means, the magnitude depending on the variability and amount of data per region. This means regions with fewer states or regions with extreme results are shrunk more strongly toward the population mean. We found that this was a good balance between ignoring regional differences (complete pooling) and treating regions independently (no pooling). Therefore, our regional parameters are modeled as draws from our population-level distribution:

```
      alpha_region = pm.Normal('alpha_region', mu=mu_alpha,
            sigma=sigma_alpha, shape=n_regions)
beta_pm25_region = pm.Normal('beta_pm25_region', mu=mu_beta_pm25,
                                    sigma=sigma_beta_pm25,
                  shape=n_regions)
```

For our priors, we chose weakly informative priors to allow the data to dominate, but still incorporating knowledge about potential ranges. For the population-level intercept, we chose *mu_alpha* under a normal distribution (`mu_alpha ~ Normal(6.0, 2.0)`) which centers at the national COPD estimates (4-8%). For the PM2.5 population effect, we also used a normal distribution (`mu_beta_pm25 ~ Normal(0.5, 1.0)`); our prior belief is that air pollution increases COPD risk but it allows for a wide uncertainty. For the effect of gender gap, we used a normal distribution (`mu_beta_gender = pm.Normal('mu_beta_gender', mu=0, sigma=1.0)`), another weakly informative prior. We assume that the gender gap is constant across regions; from our EDA we believe that this is a reasonable assumption. For our standard deviation parameters, we chose half-normal priors with sigma of 1.0 because it ensures a positive value but staying

uninformative about the magnitude. In general, our priors ensure sufficient posterior distributions whilst having minimal influence on our final estimates.

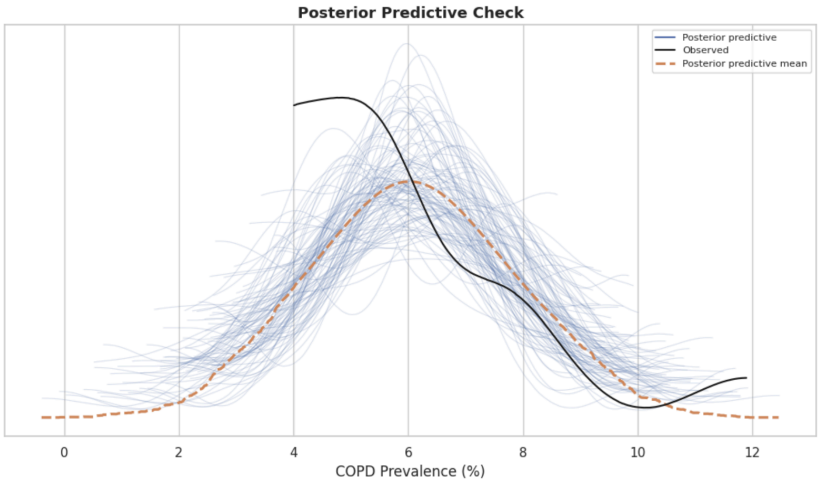
In PyMC, we implemented our model to approximate the posterior.

```
trace = pm.sample(2000, tune=1000, chains=2, return_inferencedata=True,
                  random_seed=42, target_accept=0.95,
                  idata_kwargs={"log_likelihood": True})
```

We used the NUTS sampler with 2 chains, 1,000 tune and 2,000 draws each, resulting in 6,000 total draws. We fixed the random seed at 42 for reproducibility.

Results

With some divergences noted, both chains converged to broadly similar posterior distributions, and sampling was sufficient for drawing approximate inferences about posterior means and credible intervals for most population-level parameters. Diagnostics indicate minor convergence issues, particularly for the global intercept and regional variance parameters (R-hat ≈ 1.02 and low tail effective sample sizes), suggesting that these estimates should be interpreted with caution. Posterior predictive checks (PPCs) were used to assess model fit. We generated 100 datasets from the posterior predictive distribution and compared them to the observed COPD prevalence data. The posterior predictive mean (orange dashed line) aligned reasonably well with the observed data (black line), indicating that the model captures the overall central tendency and spread of the data reasonably well. Taken together, these checks suggest that the model assumptions are acceptable for exploratory inference, though uncertainty remains for some parameters.



Our global intercept, μ_a, representing the overall age-adjusted COPD prevalence across all regions after accounting for PM2.5 exposure and gender, was estimated at 6.05%, with a 94% highest density interval (HDI) of [4.28%, 7.29%]. This estimate aligns closely with the national COPD prevalence of approximately 6% reported by the CDC, suggesting that the model is reasonably calibrated at the population level, albeit with moderate uncertainty.

Parameter	Mean	SD	95% CI lower	95% CI upper
Global intercept	6.054%	0.688	4.282	7.285
PM2.5 effect	0.303	0.330	-0.269	0.983
Gender gap effect	0.060	0.225	-0.356	0.471
Regional variation	0.976	0.465		

For the population-level PM2.5 effect, we interpret the estimate as follows: for every one standard deviation increase in PM2.5 concentration, COPD prevalence is estimated to increase by approximately 0.3 percentage points on average across regions. While the posterior mean is positive, the estimate is accompanied by substantial uncertainty. The 94% HDI includes zero, indicating that the PM2.5 effect is not robustly distinguishable from zero at the population level. This does not imply that PM2.5 has no relationship with COPD prevalence; rather, it suggests that the effect may vary across regions or be confounded by unmeasured variables not included in the model.

When examining regional effects, there is clear evidence of heterogeneity across regions. Baseline COPD prevalence varied meaningfully between regions, with posterior regional intercepts spanning approximately 5.3% to 6.9%, corresponding to a difference of roughly 1.6 percentage points, or about 27% relative variation. These differences likely reflect broader geographic patterns such as historical smoking rates, socioeconomic conditions, population age structure, and access to healthcare. The use of partial pooling shrinks regional estimates toward the population mean (~6%) while preserving meaningful differences, further supporting the appropriateness of a hierarchical modeling approach for this analysis.

In summary, the Bayesian hierarchical model reveals that COPD prevalence varies substantially across U.S. regions, with regional context playing a dominant role in explaining variation. PM2.5 exposure shows a small positive but highly uncertain association with COPD prevalence, suggesting that any effect is likely heterogeneous across regions or confounded by omitted variables. The gender gap coefficient is small and highly uncertain, indicating that while gender disparities in COPD prevalence exist, differences in the magnitude of these gaps across states do not meaningfully predict overall state-level COPD prevalence.

To assess the robustness of these findings, we conducted sensitivity analyses by varying the prior on the regional PM2.5 slope standard deviation (σ_{β}_PM2.5). Specifically, we examined three specifications: a base model with σ_{β}_PM2.5 ~ HalfNormal(0.5), a more constrained model with σ_{β}_PM2.5 ~ HalfNormal(0.25), and a more permissive model with σ_{β}_PM2.5 ~ HalfNormal(1.0). Across all three specifications, the population-level PM2.5 effect remained in the 0.25 - 0.35 range, with credible intervals consistently including zero. While the width of the credible intervals varied, being narrower under the more constrained prior and wider under the more permissive prior, the qualitative conclusion remained unchanged: the PM2.5 effect is positive in expectation but not robustly identified at the population level.

We also examined sensitivity to the grouping structure by reestimating the model using state level fixed effects rather than regional random effects. Under this specification, the PM2.5 coefficient became even less certain, with substantially larger posterior standard deviation, reflecting insufficient effective sample sizes at the state level. This result further supports the use of regional hierarchical structure to balance flexibility with statistical stability.

Overall, the effect of PM2.5 is not robust at the population level, and we cannot conclusively identify a uniform population-average effect using the available data. However, the consistently positive point estimates and the presence of meaningful regional heterogeneity suggest that PM2.5 effects may be region-dependent rather than uniformly absent.

To reduce uncertainty and improve estimation in future work, additional data and refinements would be valuable. In particular, future analyses should incorporate state-level smoking rates as confounders, increase effective sample size by incorporating longer time series, include county level data in high data states, and use historical PM2.5 exposure measures (e.g. 10-20 year averages) that better align with COPD development timescales.

We also wanted to note that we obtained much better convergence when running the notebook in Google Colab. This was interesting, and we explored to see why; one strong possibility is that the packages are different in each environment, raising another possible limitation to this modeling.

Discussion

Performing a comparison between the hierarchical model and pooled model, we found that the hierarchical model had a better predictive accuracy (lower WAIC, Widely Applicable Information Criterion) with a moderate but meaningful WAIC difference of 4.71. Additionally, the added complexity of the model is justified because of its better predictive performance.

However, our methods did highlight some limitations. Firstly, we only used 2020 data which prevents inference about PM2.5 effects and temporal trends. For example, COPD could be a disease that develops over time, and not necessarily reflected in the current year of air quality concentrations. Additionally, we have to acknowledge that 2020 values might not represent typical conditions due to the COVID-19 pandemic, so historical data would likely have been more appropriate. Another limitation was that we did not include some variables that might have been important. For example, smoking prevalence is a big risk factor for COPD. We could argue that this could be captured by the regional effects, but it would likely have been useful to have state-level smoking statistics. Other similar factors could be socioeconomic factors, maybe regional-level, which could be correlated with PM2.5 and COPD.

We did find that we had some divergence during sampling. This often happens when we have limited groups and small group sizes. We only had 4 regions, some of which were relatively small which could contribute to these divergences. Although we did obtain good R-hat values, we could improve robustness using longer MCMC runs, or testing out alternative parameters.

Connecting our research question to Lui et al. 's *Trends in the Prevalence of Chronic Obstructive Pulmonary Disease Among Adults Aged >= 18 years-US 2011-2021*, some of their findings does validate our results. They saw a constant COPD prevalence throughout the years but wanted to delve further into other factors. Firstly, they also found a geographical variation. Our finding of substantial regional variation in COPD prevalence is consistent with recent CDC surveillance data showing state-level prevalence ranging from 3.0% to 11.8% (Liu et al., 2023), supporting the use of hierarchical modeling approaches that account for geographic clustering. Another interesting similarity was firstly that their COPD data was collected from surveys, and highlighting similar limitations in terms of response bias. They mentioned the possibility of 2020 influencing their results. For them it was in terms of reporting; they found that there were no

differences between 2019, 2020 and 2021 to influence COPD prevalence, potentially making our model more generalized. There were two major differences for this article: they included urban and rural information and looked at smoking data. They found that "COPD prevalence increased with increasing age, decreasing education level, and decreasing urbanicity" (Liu et al., 2023). Because we used state-level and regional-level COPD, we did not capture this urbanicity factor. Smoking data was a big focus in their report, reinforcing that this is an important factor when exploring COPD, due to smoking levels, quitting rate and the age of the area in relation to smoking prevalence.

Research Question 2

Methods

To explore the causal relationship between PM2.5 concentrations and COPD prevalence, our outcome variable is the age-adjusted COPD prevalence per state. The treatment we used is 'elevated PM2.5 concentrations' in which we made a binary variable of whether the PM2.5 concentration predictions for the year 2020 above the 75th percentile of our state data. Above the 75th percentile represents elevated levels.

The confounding variables that we considered were state smoking percentages, state urban household percentages, and the median household income of the state. The urban household percentage and median household income initially may not seem like confounders, however you could argue that both of these variables can have a causal effect on both outcome and treatment variables. When median household income is higher, there might be more pollution because of increased household car usage, increasing PM2.5 concentrations. They might also have better access to healthcare, which could directly affect COPD prevalence. When more households are in an urban area, PM2.5 concentrations could be higher because of a denser population. These households could also have impacts on healthcare access and COPD diagnoses. We did not find any colliders in the dataset we used.

We decided to use outcome regression using Ordinary Least Squares. To adjust for confounders, we fitted our model with our treatment variable along with our confounding variables. The regression coefficient can be interpreted as the causal effect if we have accounted for all confounders. This method also requires us to assume unconfoundedness given our set of variables, and that the new linear model correctly describes the interaction between the variables. We do not believe that the unconfoundedness assumption holds, because there are many different variables we have not included in our model that act as confounders.

Results

Using the variables stated above, we first performed OLS with no changes. We found that we were getting extremely high condition number values, an indicator of strong multicollinearity

or other problems with our variables. We applied the standard scaler to our confounding variables, and we found that this condition number reduced and we no longer got this warning, implying that the issue was likely due to the different scales of our confounders. As above, our assumptions were that the unconfoundedness assumption holds (i.e. we have accounted for all confounders), the model is linear in its parameters, we have no perfect multicollinearity, and that each error term is independent of each other.

Using our results, our coefficient of interest or estimated treatment effect was 0.533 with a standard error of 0.343. This can be interpreted as states with PM2.5 concentrations above the 75th percentile of statewide PM2.5 concentrations have 0.53 percentage points higher COPD prevalence.

OLS Regression Results

Dep. Variable:	COPD prevalence	R-squared:	0.694
Model:	OLS	Adj. R-squared:	0.667
Method:	Least Squares	F-statistic:	24.98
Date:	Tue, 16 Dec 2025	Prob (F-statistic):	7.72e-11
Time:	04:38:43	Log-Likelihood:	-66.578
No. Observations:	49	AIC:	143.2
Df Residuals:	44	BIC:	152.6
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	5.9341	0.169	35.204	0.000	5.594	6.274
Smoking pct	1.5118	0.282	5.353	0.000	0.943	2.081
Urban pct	0.2177	0.193	1.128	0.265	-0.171	0.607
Median Household Income	-0.0151	0.271	-0.056	0.956	-0.562	0.531
above 75th percentile	0.5329	0.343	1.555	0.127	-0.158	1.223

Omnibus:	8.516	Durbin-Watson:	2.028
Prob(Omnibus):	0.014	Jarque-Bera (JB):	8.285
Skew:	0.723	Prob(JB):	0.0159
Kurtosis:	4.403	Cond. No.	4.50

The uncertainty of this is quite wide, ranging from -0.16 to 1.22 percentage points; the interval ranging from negative to positive is not a promising result, providing evidence against a causal relationship. When we inspect the p-value (0.127), we can conclude this would not be statistically significant at the 90% confidence level. We therefore cannot say that PM2.5 levels above the 75th percentile causes COPD prevalence to change.

Discussion

For this research question, one of our main limitations is the confounding variables that have not been included in this model. For a research question exploring statewide COPD prevalence and PM2.5 concentrations, there are many different directions to go such as climate, socioeconomic, geographical and demographic-wise that you could find potential confounders. It was difficult finding the correct datasets that we could join with our primary data to use as confounders, especially because we used a one year dataset (2020). We would have found more data on confounders and potentially instrumental variables extremely useful because we would be closer to achieving the assumptions we stated in the sections above. This brings us to another

limitation, which is the amount of data we had. We chose to use the year 2020 because it was the year in which there was the most overlap between datasets. However, this does not take into account that COPD is likely a disease that develops over time, and the current data might not correctly reflect the relationship. Another part of our model we could have improved on was how we identify a state's elevated PM2.5 concentration. We classified this compared to the other states, but when comparing these to the US Environmental Protection Agency standard, only a few states would show elevated concentrations. Our classification was used so that we had more of a balanced set of groups. Comparing to the standard might have been more meaningful.

Looking at the results of our model, we can conclude that there is not a causal relationship between COPD prevalence and PM2.5 concentrations being above the 75th percentile. We can say this with quite strong confidence because of our obtained p-value and confidence interval range. As stated above, our coefficient was not statistically significant at the 90% confidence level and our confidence interval includes zero, we are uncertain of any treatment effect. It is important to note that if more confounders were accounted for and if we used a more rigorous approach such as inverse propensity weighting or matching, there could possibly be a causal relationship between our variables.

Our results were different to Park et al.'s study mentioned in the prior work section. They did find an association between 10 $\mu g/m^3$ increase in PM2.5 concentrations with an 18% increased incidence of COPD (Park et al., 2021). Their results were also statistically significant, but did not classify explicitly a causal relationship because they were only looking at the relationship. Another difference was the years in which they were analyzing. All of the studies they had access to had follow-up studies; this could be a reason why they achieved meaningful results compared to our methods because they were able to account for how COPD develops and not just one year.

Conclusions

Our analysis across both research questions suggests that COPD prevalence in the United States is highly affected by a significant regional and demographic heterogeneity, with very little uncertainty explained by the PM2.5 levels. By using CDC's indicators of age-adjusted COPD prevalence, PM2.5 predictions aggregated up to state level, and demographic data, we built a Bayesian hierarchical model to analyze cross-state variations (RQ1), and an OLS regression to explore causality from elevated PM2.5 to COPD prevalence (RQ2).

For RQ1, the hierarchical model we posed revealed clear regional clustering with the baseline COPD prevalence varying by over 1.5 percentage points across different regions (with global intercept close to 6%). At the same time we observed a positive yet uncertain population-level association (around a 0.3 percentage point increase in prevalence per SD increase, with a credible interval including zero), and gender gaps didn't have a statistically significant effect in the prediction of state level prevalence. The PPCs and WAIC analyses resulted in favor of the hierarchical model instead of the pooled alternative, reinforcing the idea that regional structure is statistically significant.

For RQ2, we defined treatment as having PM2.5 levels above the 75th percentile of the state's distribution in 2020 and adjusted it for smoking prevalence, share of urban households from the total, and median household income using OLS regression. The estimated treatment effect for positive (high-PM2.5) states was a 0.533% higher COPD prevalence, but the confidence interval is (-0.16, 1.22) which crosses zero, and the p-value is 0.127 which is way higher than conventional thresholds. With this, we cannot claim a statistically significant causal effect of elevated PM2.5 on COPD prevalence at the state level, in contrast to other prior individual level work that found significant association between long term PM2.5 exposure and COPD incidence levels.

Several limitations appear for both our questions. The main one being that COPD and pollution exposure evolve over decades but our analysis relies only on 2020 cross-sectional data, limiting our ability to capture exposure windows important for inferring causality. We also lack key confounders such as occupation based exposure, healthcare access, or even more detailed smoking histories, and we aggregated COPD and PM2.5 to a state level, which can cover intra-state disparities. A key question for a pulmonary epidemiologist could be: over what long term window is PM2.5 exposure linked to start having effects on the COPD risk, or even in which phase of life is it considered to have harsher effects?

Our findings are only moderately robust to modeling variations. In RQ1, different assumptions of how regional differences and PM2.5 impact modify the uncertainty of the coefficients but they would still preserve the strong regional structures found. In RQ2 , changing treatment thresholds, more comprehensive confounder sets, or other causal techniques could also adjust the impact closer or further from zero. Therefore the results should be seen as indicative patterns under specific assumptions, rather than as absolute causal claims, and its generalizability is limited to state level trends in 2020 rather than individual risk levels.

Future research should integrate multi year PM2.5 exposure histories, individual level BRFSS microdata, and other strong confounders not accessed in this study. Another element to build on is to use stronger causal methodologies and perhaps a more complex hierarchical framework like nesting counties inside states and regions. The study does not demonstrate significant causal relationship between elevated PM2.5 levels and COPD prevalence. However, the descriptive clustering of COPD, the slight increasing effects predicted on pollution, and extensive epidemiological literature all support the investment in air quality regulations and COPD prevention. In reality this could mean putting screening, resources to help people quit smoking, and clean air treatments prioritizing the areas where they are the most common. It also means making sure that the cost of managing pollution doesn't end up falling too heavily on already struggling communities.

Appendix

Liu, Y., Carlson, S. A., Watson, K. B., Xu, F., & Greenlund, K. J. (2023). Trends in the prevalence of chronic obstructive pulmonary disease among adults aged ≥18 years—United States, 2011–2021. *Morbidity and Mortality Weekly Report*, 72(46), 1251-1256. https://doi.org/10.15585/mmwr.mm7246a1

Park, J., Kim, H.-J., Lee, C.-H., Lee, C. H., & Lee, H. W. (2021). Impact of long-term exposure to ambient air pollution on the incidence of chronic obstructive pulmonary disease: A systematic review and meta-analysis. *Environmental Research*, *194*, Article 110703. https://doi.org/10.1016/j.envres.2020.110703