# Understanding the Factors Influencing COPD Prevalence in the United States

Narek Tahmazyan

ECON 409: Introduction to Econometrics

Professor Dennis Halcoussis

December 7, 2023

## Introduction:

In our rapidly evolving and urbanizing world, understanding the impact of environmental changes on our health becomes crucial. With problems such as global warming changing the natural environment and cities becoming more polluted, respiratory diseases are starting to become a more prevalent issue. This paper delves into the dynamic relationship between chronic obstructive pulmonary disease and a range of environmental and socioeconomic variables, shedding light on their collective influence on the prevalence of COPD among adults. By examining factors such as cigarette smoking, physical inactivity, and a few other environmental variables, this study aims to unravel the complex web of connections that contribute to the prevalence of COPD.

COPD, a term encompassing chronic bronchitis and emphysema, has emerged as a critical public health concern, with its prevalence steadily rising worldwide (Adeloye et al., 2015). This multifaceted disease not only places a substantial burden on healthcare systems but also severely impacts individuals' quality of life. To address this growing health crisis, it is crucial to examine the potential environmental and socioeconomic drivers of COPD. Many of the articles and research papers that will be discussed later are studies conducted in different parts of the world, including populations from a variety of socioeconomic backgrounds.

## Literature Review:

According to the study conducted by M. Bartal in 2005, active smoking was identified as a primary risk factor contributing to the development and exacerbation of COPD. Furthermore, it brings attention to how even indirect exposure to tobacco smoke, especially during childhood, can diminish pulmonary function while increasing susceptibility to developing COPD later on in their lifetime. For example, the study notes that secondhand smoking in children can lead to a reduction in FEV1 (forced expiratory volume in 1

second) of about 5–7% by the age of 14, primarily due to in utero exposure (Bartal, 2005). Moreover, the study emphasizes the critical role of quitting smoking in slowing the progression of COPD, presenting evidence that it can significantly reverse the decline in lung function, especially if done early. For instance, study results from Fletcher et al. (1977) showed a 10-year survival rate of 80% in ex-smokers with COPD, compared to only 50% in those who continued to smoke.

In a multi-center pilot study, Troosters et al. (2010) investigated the relationship between physical inactivity and COPD. The study involved 70 COPD patients and 30 healthy controls and found a significant reduction in physical activity among COPD patients, starting as early as GOLD-stage II (COPD classification characterized as “moderate”). Specifically, the number of daily steps and time spent in moderate-intensity activities were noticeably lower in COPD patients compared to the control group. This reduction in physical activity was proportional to the severity of the disease, which shows a clear inverse relationship between COPD progression and physical activity levels (Troosters et al., 2010).

Another study conducted in South Korea by Lee et al. (2019) analyzed the relationship between poverty and COPD. The research, based on data from 3223 individuals in South Korea, revealed that 25.8%, or 832 participants, lived below the relative poverty line, and 11.9% of them had COPD. Notably, the prevalence of COPD was higher among those living below the poverty line compared to those above it. This study found a 1.4-fold increased risk of developing COPD among individuals living in poverty. Furthermore, elderly individuals living below the poverty line were 1.5 times more likely to develop COPD than those living above it (Lee et al., 2019).

The study, conducted by Gou et al. (2023), analyzed data from 313,013 COPD deaths in Chongqing between 2012 and 2020, aiming to analyze the relationship between the availability of green spaces and COPD mortality. It utilized Fractional Vegetation Cover (FVC), derived from the Normalized Difference Vegetation Index (NDVI) from satellite imagery, to define green spaces. The study discovered that 63% of regions in Chongqing showed a positive correlation between green spaces and COPD mortality, while 37% exhibited a negative correlation. The interpretive power of the FVC factor on the spatial distribution

of COPD mortality was found to be 0.08 (Gou et al., 2023). These findings suggest that the impact of green spaces on COPD mortality is spatially varied and potentially influenced by factors beyond the presence of greenery alone, such as access to medical resources and regional socioeconomic conditions.

Lastly, a study called "Is Exposure to Biomass Smoke the Biggest Risk Factor for COPD Globally?" reveals a compelling link between biomass smoke exposure and COPD. Biomass smoke, predominantly from burning wood, animal dung, and crop residues, is identified as a critical contributor to indoor air pollution, particularly in rural and developing regions. Hu et al. (2010) find that exposure to biomass smoke nearly doubles the risk of developing COPD, a risk level comparable to that of cigarette smoking. This association is particularly noteworthy given the extensive global reliance on biomass fuel, implying that biomass smoke might be one of the most significant worldwide risk factors for COPD (Capistrano et al., 2017).

**Methodology and Data:**

For this study, I am going to be using cross-sectional data. Specifically, the cross-sectional dimension is represented by the different states. Since the data for each variable came from different sources, some sources included the District of Columbia as one of the "states," but for simplicity and continuity of the data, we are going to drop DC from all of the datasets. Combining 50 states and 6 variables, we get 300 observations; therefore, the sample size is 300. Degrees of freedom (n-k-1): 300 - 6 - 1 = 293 which is a generous enough number considering I am not using time-series models that can quickly deplete degrees of freedom.

There were some challenges when it came to finding relevant data. For example, my initial idea was to add a variable for average air quality in a year for each of the 50 states, but to my surprise, this type of data wasn't readily available, so I decided to instead use the variable called FIRES, which will be described later in the variables section of the paper.

There was also an issue with getting the data from the same year. The aim was to get the most recent and relevant data, and hopefully, the gap between the years for each dataset would be minimal. Three out of six variables have data from 2023, one from 2021, and two from 2020. Refer to Appendix 7 for the links to the data sources.

The model will look like this:

$$\mathbf{COPD = B0 + B1SMOKE + B2PHYSINAC + B3GREEN + B4FIRES + B5POVERTY + e}$$

## Variables:

**COPD:** Prevalence of chronic obstructive pulmonary disease (COPD) among adults in each state, measured as a percentage of the state's adult population.

**SMOKE**: Rate of adult cigarette smoking in each state, measured as the percentage of adults who are current smokers. Smoking is the most significant risk factor for COPD, with the majority of cases being directly attributable to tobacco use. The inclusion of this variable is based on a plethora of studies that have consistently demonstrated the damaging impact of cigarette smoke on lung tissue. Beyond the primary effects of smoking, secondhand smoke exposure during childhood and even in utero can significantly reduce pulmonary function and increase susceptibility to developing COPD later in life.

**PHYSINAC:** Prevalence of physical inactivity among adults in each state, measured as the percentage of adults who do not meet the recommended physical activity threshold. Regular physical activity is known to enhance respiratory function and overall health, reducing the risk and severity of COPD. By including this variable, the model acknowledges the role of lifestyle factors in respiratory health. The assumption is that higher rates of inactivity correlate with greater COPD prevalence.

**GREEN:** Coverage of state and national parks in each state. It is measured as the percentage of the state's total land area that is covered by parks. This variable serves as a proxy for access to natural environments, which are believed to promote pulmonary health through cleaner air and opportunities for physical activity. However, the relationship between green spaces and COPD is complex, with some studies suggesting that proximity to greenery can improve respiratory symptoms, while others indicate potential allergen exposures.

**FIRES:** The impact of forest fires in each state. It is measured as the percentage of the state's total forest acreage that was burned. Forest fires release large quantities of particulates and pollutants into the air, exacerbating respiratory conditions. This variable is particularly relevant given increasing concerns about the health impacts of wildfire smoke, which can travel vast distances and affect populations far from the fire source. Furthermore, the relevance of this variable is underscored by studies examining the impact of biomass smoke on respiratory health.

**POVERTY:** Economic hardship in each state. It is measured as the percentage of the state's total population living below the poverty line. The inclusion of this factor is based on the understanding that lower socio-economic status is associated with a range of risk factors for COPD, including increased tobacco use, occupational hazards, and limited access to healthcare.

### OLS Results:

Before interpreting the regression results, it is important to mention that I selected the robust standard errors option before running the regression because I am using cross-sectional data. Therefore, it is important to have heteroskedasticity-robust standard errors. For more detailed information about the OLS result, see Appendix 2.

$$\mathbf{COPD} = -3.20297 + 0.300041\mathbf{SMOKE} + 0.240282\mathbf{PHYSINAC} - 0.0586912\mathbf{GREEN} + 0.902253\mathbf{FIRES} - 0.117232\mathbf{POVERTY} + e$$

**SMOKE:** The positive and statistically significant coefficient indicates that for every 1% increase in the smoking rate, COPD prevalence increases by 0.300041 percentage points, holding other variables constant. This underscores the well-established link between smoking and COPD.

**PHYSINAC**: Ceteris paribus, for each percentage point increase in the physical inactivity rate among adults, the COPD rate increases by 0.2403 percent, keeping other variables constant. This finding aligns with existing literature on the impact of exercise on respiratory health.

**GREEN:** The coefficient for green space/park coverage is negative; however, this relationship is not statistically significant (p-value: 0.2847), indicating that within the context of this model, the effect of the availability of state and national parks on COPD prevalence is inconclusive.

**FIRES:** For every 1% increase in the proportion of forest acreage burned, there's an associated increase of about 0.902 percentage points in COPD prevalence, all else being equal. This result is the most significant one out of all the variables that we have, which highlights the potential impact of air quality worsening due to wildfires on respiratory health.

**POVERTY:** The coefficient for poverty is -0.117232 but not statistically significant (p-value: 0.1065). The inverse relationship between poverty rates and COPD prevalence also looks irrational, but because of the p-value, we can ignore this result.

**Model Summary:** The independent variables explain a significant portion of the variability in COPD prevalence, which is explained by the adjusted R-squared value of 0.740083. The OLS result indicates that first-hand smoking rates, physical inactivity, and the impact of forest fires are key variables in predicting COPD prevalence.

### OLS Violations:

Given that I am working with cross-sectional data, the main concern when it comes to OLS violations is heteroskedasticity. To test for it, I am using White's test. The result is a p-value of 0.023566, which indicates the presence of heteroskedasticity (for further details, refer to Appendix 3). Since all the

variables are in percentage terms, the model is already designed in a way to account for the difference in observation sizes. Also, we cannot use the log transformation function either; I will leave the model as it is since the original model already has a robust standard error option selected, which means the model already has unbiased standard errors and adjusted p-values.

Next, I analyze the correlation coefficients for each variable using the correlation matrix function on Gretl (as shown in Appendix 4) in order to find any variables that could exhibit strong correlations with one another. There is a strong positive correlation between COPD and PHYSINAC (r-value of 0.7780) and a moderate negative correlation between COPD and GREEN (r-value of -0.3482). None of the correlation values are above 0.8, which could indicate a strong positive linear relationship between the variables.

Lastly, in order to test for multicollinearity, I employ the variable inflation factor. The results greater than 4 indicate multicollinearity. The VIF results (Appendix 5) show that the largest value is 2.484 for PHYSINAC and the smallest one is 1.251 for GREEN, indicating no multicollinearity.

**Robustness:**

It is important to check for the robustness of the model, making sure it can handle the unexpected and adapt to new data. In order to test the robustness, I removed two of the weakest performing variables, which are GREEN and POVERTY.

The coefficients of the variables have slightly changed in the new model, but nevertheless, the remaining variables (SMOKE, PHYSINAC, FIRES) maintained their significance, as indicated by their p-values (all $p<0.05$). The R-squared value of 0.743036 indicates that approximately 74.3% of the variation in COPD prevalence can be explained using the model. This suggested that these variables are robust predictors of COPD prevalence, even when controlling for potential heteroskedasticity in the data. For further analysis of the results, refer to Appendix 6.

## Conclusion:

The key findings revealed that smoking and physical inactivity are significant contributors to COPD prevalence, aligning with existing literature on the harmful effects of these factors on respiratory health. The poverty and park/green space accessibility variables were not statistically significant. Although one of the environmental variables, FIRES, had the biggest coefficient, showing that poor air quality is one of the major contributors to COPD prevalence. However, the regression results showed that lifestyle choices also have a huge impact, and specifically, those are the variables that the individual can control in order to reduce the risk of COPD.

Interestingly, the robustness check of the model, which involved removing the weaker variables, further emphasized the significance of smoking, physical inactivity, and forest fires as robust predictors of COPD prevalence. This refinement of the model not only reinforced the initial findings but also enhanced the reliability of our conclusions.

However, the presence of heteroskedasticity in the model, as indicated by White's test, necessitated the use of robust standard errors to ensure the stability of our results and mitigate the effects of heteroskedasticity.

## Work Cited:

Bartal, M. “COPD and tobacco smoke.” *Monaldi archives for chest disease = Archivio Monaldi per le malattie del torace* vol. 63,4 (2005): 213-25. doi:10.4081/monaldi.2005.623

Troosters, Thierry et al. “Physical inactivity in patients with COPD, a controlled multi-center pilot-study.” *Respiratory medicine* vol. 104,7 (2010): 1005-11. doi:10.1016/j.rmed.2010.01.012

Lee, Young Seok et al. “The association between living below the relative poverty line and the prevalence of chronic obstructive pulmonary disease.” *Journal of thoracic disease* vol. 11,2 (2019): 427-437. doi.org:10.21037/jtd.2019.01.40

Gou, A., Tan, G., Ding, X. *et al.* Spatial association between green space and COPD mortality: a township-level ecological study in Chongqing, China. *BMC Pulm Med* 23, 89 (2023): doi:10.1186/s12890-023-02359-x

Capistrano, Sarah J et al. “Evidence of Biomass Smoke Exposure as a Causative Factor for the Development of COPD.” *Toxics* vol. 5,4 36. 1 Dec. 2017, doi:10.3390/toxics5040036

Adeloye, Davies et al. “Global and regional estimates of COPD prevalence: Systematic review and meta-analysis.” *Journal of global health* vol. 5,2 (2015): 020415. doi:10.7189/jogh.05.020415

**Appendix:**

Appendix 1: Summary Statistics

| | Mean | Median | Minimum | Maximum |
|---|---|---|---|---|
| COPD | 6.0180 | 5.6500 | 3.2000 | 11.900 |
| PHYSINAC | 24.932 | 24.650 | 17.700 | 33.200 |
| POVERTY | 12.486 | 12.050 | 7.4000 | 19.500 |
| GREEN | 1.9520 | 1.0000 | 0.060000 | 9.4100 |
| FIRES | 0.20300 | 0.050000 | 0.00000 | 2.2000 |
| SMOKE | 15.416 | 15.650 | 8.2000 | 22.600 |

| | Std. Dev. | C.V. | Skewness | Ex. kurtosis |
|---|---|---|---|---|
| COPD | 1.7509 | 0.29094 | 1.2996 | 1.9197 |
| PHYSINAC | 3.6924 | 0.14810 | 0.23925 | -0.41123 |
| POVERTY | 2.6016 | 0.20836 | 0.82995 | 0.42889 |
| GREEN | 2.4625 | 1.2615 | 1.7961 | 2.1255 |
| FIRES | 0.43535 | 2.1446 | 3.0365 | 9.0937 |
| SMOKE | 3.2958 | 0.21379 | -0.0015019 | -0.64910 |

| | 5% perc. | 95% perc. | IQ range | Missing obs. |
|---|---|---|---|---|
| COPD | 4.0550 | 9.8900 | 2.4250 | 0 |
| PHYSINAC | 18.310 | 31.730 | 5.3000 | 0 |
| POVERTY | 9.0300 | 18.375 | 3.4250 | 0 |
| GREEN | 0.086500 | 8.2145 | 1.6950 | 0 |
| FIRES | 0.00000 | 1.4020 | 0.10000 | 0 |
| SMOKE | 9.9450 | 20.905 | 4.9750 | 0 |

Appendix 2: OLS Results

Model 2: OLS, using observations 1-50
Dependent variable: COPD
Heteroskedasticity-robust standard errors, variant HC1

| | coefficient | std. error | t-ratio | p-value | |
|---|---|---|---|---|---|
| const | -3.20297 | 1.29931 | -2.465 | 0.0177 | ** |
| SMOKE | 0.300041 | 0.0726110 | 4.132 | 0.0002 | *** |
| PHYSINAC | 0.240282 | 0.0510663 | 4.705 | 2.54e-05 | *** |
| GREEN | -0.0586912 | 0.0541958 | -1.083 | 0.2847 | |
| FIRES | 0.902253 | 0.313674 | 2.876 | 0.0062 | *** |
| POVERTY | -0.117232 | 0.0711465 | -1.648 | 0.1065 | |

| | | | |
|---|---|---|---|
| Mean dependent var | 6.018000 | S.D. dependent var | 1.750882 |
| Sum squared resid | 35.05913 | S.E. of regression | 0.892636 |
| R-squared | 0.766605 | Adjusted R-squared | 0.740083 |
| F(5, 44) | 15.85367 | P-value(F) | 6.47e-09 |
| Log-likelihood | -62.07226 | Akaike criterion | 136.1445 |
| Schwarz criterion | 147.6166 | Hannan-Quinn | 140.5132 |

Excluding the constant, p-value was highest for variable 5 (GREEN)

Appendix 3: White's test

```
White's test for heteroskedasticity
OLS, using observations 1-50
Dependent variable: uhat^2

               coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const          14.8266       5.57738     2.658     0.0126  **
  SMOKE          -2.17721      0.636325   -3.422     0.0019  ***
  PHYSINAC        0.134233     0.642666    0.2089    0.8360
  GREEN          -1.54390      1.08504    -1.423     0.1654
  FIRES          -6.40251     11.1587     -0.5738    0.5705
  POVERTY         0.341999     0.641260    0.5333    0.5979
  sq_SMOKE        0.0577578    0.0310334   1.861     0.0729  *
  X2_X3           0.0564451    0.0527123   1.071     0.2931
  X2_X4           0.0472183    0.0332008   1.422     0.1656
  X2_X5           0.809897     0.617388    1.312     0.1999
  X2_X6          -0.0829685    0.0429349  -1.932     0.0631  *
  sq_PHYSINAC    -0.0284729    0.0260689  -1.092     0.2837
  X3_X4           0.0557983    0.0378725   1.473     0.1514
  X3_X5          -0.206635     0.272467   -0.7584    0.4543
  X3_X6           0.0206017    0.0283967   0.7255    0.4740
  sq_GREEN        0.00330848   0.0400532   0.08260   0.9347
  X4_X5           0.548176     0.279101    1.964     0.0592  *
  X4_X6          -0.0481549    0.0672774  -0.7158    0.4799
  sq_FIRES        0.132787     1.55949     0.08515   0.9327
  X5_X6          -0.0805997    0.292741   -0.2753    0.7850
  sq_POVERTY      0.0202890    0.0181288   1.119     0.2723

  Unadjusted R-squared = 0.687924

Test statistic: TR^2 = 34.396205,
with p-value = P(Chi-square(20) > 34.396205) = 0.023566
```

Appendix 4: Correlation Matrix

```
Correlation Coefficients, using the observations 1 - 50
Two-tailed critical values for n = 50: 5% 0.2787, 1% 0.3610

          COPD      PHYSINAC       POVERTY         GREEN         FIRES
        1.0000        0.7780        0.3424       -0.3482       -0.2204  COPD
                      1.0000        0.5036       -0.3422       -0.3479  PHYSINA
                                    1.0000       -0.1807        0.0804  POVERTY
                                                  1.0000        0.3586  GREEN
                                                                1.0000  FIRES
```

Appendix 5: Variance Inflation Factors

```
Variance Inflation Factors
Minimum possible value = 1.0
Values > 10.0 may indicate a collinearity problem

       SMOKE    2.291
    PHYSINAC    2.484
       GREEN    1.251
       FIRES    1.452
     POVERTY    1.534
```

Appendix 6: Robustness Check

```
Model 4: OLS, using observations 1-50
Dependent variable: COPD
Heteroskedasticity-robust standard errors, variant HC1

             coefficient   std. error   t-ratio   p-value
  ---------------------------------------------------------
  const       -3.70936     1.26410      -2.934    0.0052   ***
  SMOKE        0.290313    0.0733320     3.959    0.0003   ***
  PHYSINAC     0.205798    0.0512374     4.017    0.0002   ***
  FIRES        0.595823    0.269250      2.213    0.0319   **

Mean dependent var   6.018000   S.D. dependent var   1.750882
Sum squared resid    38.59959   S.E. of regression   0.916036
R-squared            0.743036   Adjusted R-squared   0.726277
F(3, 46)             22.98171   P-value(F)           3.07e-09
Log-likelihood      -64.47739   Akaike criterion     136.9548
Schwarz criterion    144.6029   Hannan-Quinn         139.8672
```

Appendix 7: Data Source Links

https://www.cdc.gov/copd/data-and-statistics/state-estimates.html (COPD)

https://wisevoter.com/state-rankings/smoking-rates-by-state/ (SMOKE)

https://www.cdc.gov/physicalactivity/data/inactivity-prevalence-maps/index.html (PHYSINAC)

https://www.playgroundequipment.com/us-states-ranked-by-state-and-national-park-coverage/ (GREEN)

https://wisevoter.com/state-rankings/most-forested-states/ (FIRES)

https://data.ers.usda.gov/reports.aspx?ID=17826 (POVERTY)