# Geisinger Collider Project

## Predicting COPD in pneumonia patients

Phase 2 - Spring 2016

Rebecca Barter and Shamindra Shrotriya

## Question:

**For a new patient who has been diagnosed with pneumonia, do they have Chronic Obstructive Pulmonary Disease (COPD)?**

Can incorporation of external information improve prediction?

# Chronic Obstructive Pulmonary Disease

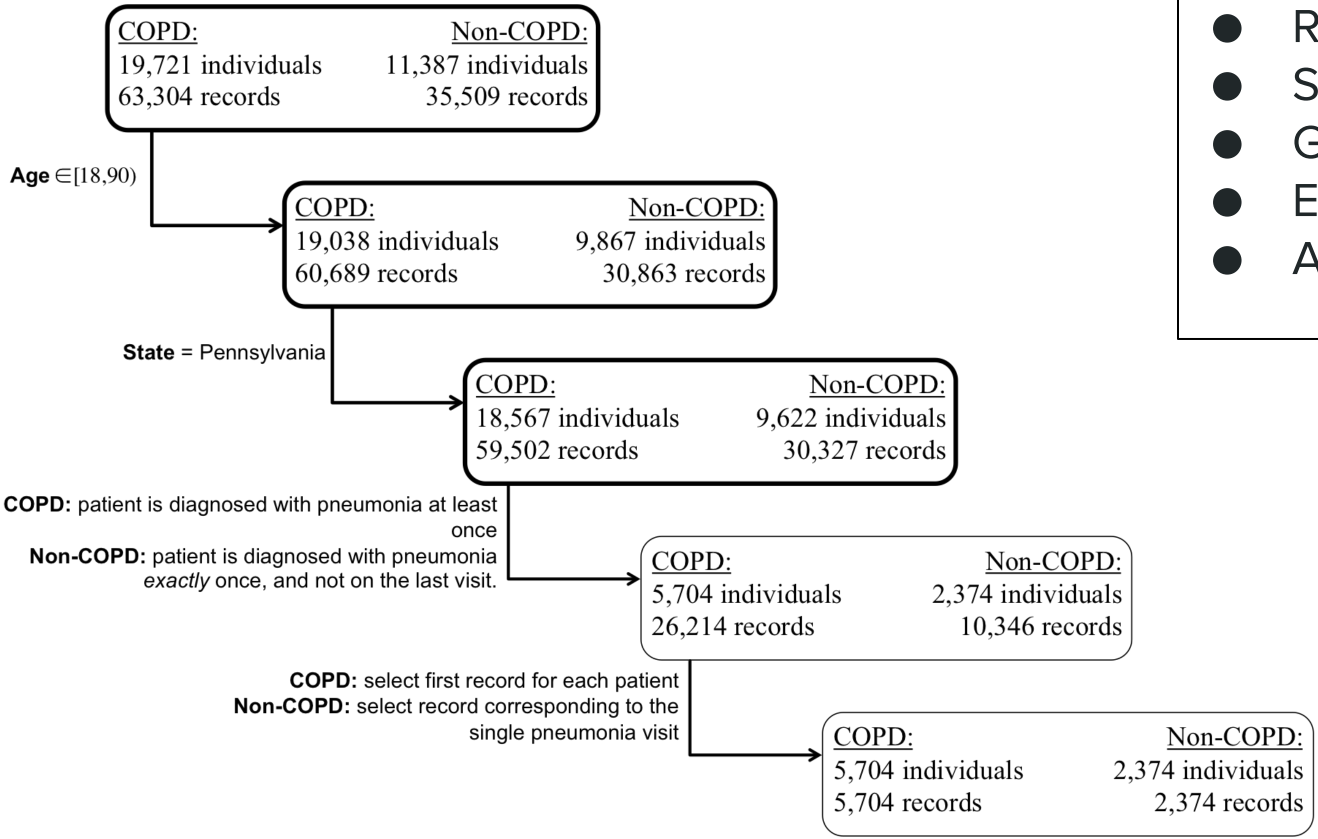
- **COPD is a major cause of mortality worldwide.**
- **Approximately 12 million adults in the U.S. having been diagnosed with COPD .**
- **A further 12 million adults in the U.S. are currently living with undiagnosed COPD.**

# Key Hypotheses: COPD Risk Factors

- **Smoking**
  - Available from Geisinger clinical data!
- **Occupational exposure to VOCs (emissions from biomass fuels)**
  - Infer from employment information provided by Geisinger!
- **Outdoor pollution**
  - Find on the internet
- **Weather**
  - Find on the internet

# The Data Collection

# Clinical Data




- Age
- Race
- Smoking
- Gender
- Employment
- Asthma

# Daily pollution data from the EPA website!!!

## Daily Summary Data

### Criteria Gases

| Year | Ozone (44201) | SO2 (42401) | CO (42101) | NO2 (42602) |
|---|---|---|---|---|
| *2015* | daily_44201_2015.zip<br>259,151 Rows<br>2,958 KB<br>As of 2015-11-27 | daily_42401_2015.zip<br>213,936 Rows<br>2,198 KB<br>As of 2015-11-27 | daily_42101_2015.zip<br>129,000 Rows<br>1,141 KB<br>As of 2015-11-27 | daily_42602_2015.zip<br>95,351 Rows<br>1,309 KB<br>As of 2015-11-27 |
| *2014* | daily_44201_2014.zip<br>391,846 Rows<br>4,389 KB<br>As of 2015-11-27 | daily_42401_2014.zip<br>324,818 Rows<br>3,277 KB<br>As of 2015-11-27 | daily_42101_2014.zip<br>215,101 Rows<br>1,826 KB<br>As of 2015-11-27 | daily_42602_2014.zip<br>148,509 Rows<br>1,991 KB<br>As of 2015-11-27 |
| *2013* | daily_44201_2013.zip<br>391,592 Rows<br>4,388 KB<br>As of 2015-11-27 | daily_42401_2013.zip<br>332,132 Rows<br>3,340 KB<br>As of 2015-11-27 | daily_42101_2013.zip<br>216,689 Rows<br>1,822 KB<br>As of 2015-06-20 | daily_42602_2013.zip<br>139,272 Rows<br>1,841 KB<br>As of 2015-11-27 |
| *2012* | daily_44201_2012.zip<br>388,718 Rows<br>4,404 KB<br>As of 2015-11-27 | daily_42401_2012.zip<br>330,112 Rows<br>3,335 KB<br>As of 2015-11-27 | daily_42101_2012.zip<br>222,504 Rows<br>1,908 KB<br>As of 2015-11-27 | daily_42602_2012.zip<br>134,777 Rows<br>1,776 KB<br>As of 2015-11-27 |
| *2011* | daily_44201_2011.zip<br>381,859 Rows | daily_42401_2011.zip<br>323,535 Rows | daily_42101_2011.zip<br>226,465 Rows | daily_42602_2011.zip<br>131,819 Rows |

- Ozone
- CO
- SO2
- NO2
- PM10
- PM2.5
- Arsenic
- Lead
- NO
- CS2

# Daily weather data from PSU Climatologist website!!!

- Temperature
- Pressure
- Humidity

## PASC IDA Data Page

Select a network: FAA Daily
Select a display option: List Map

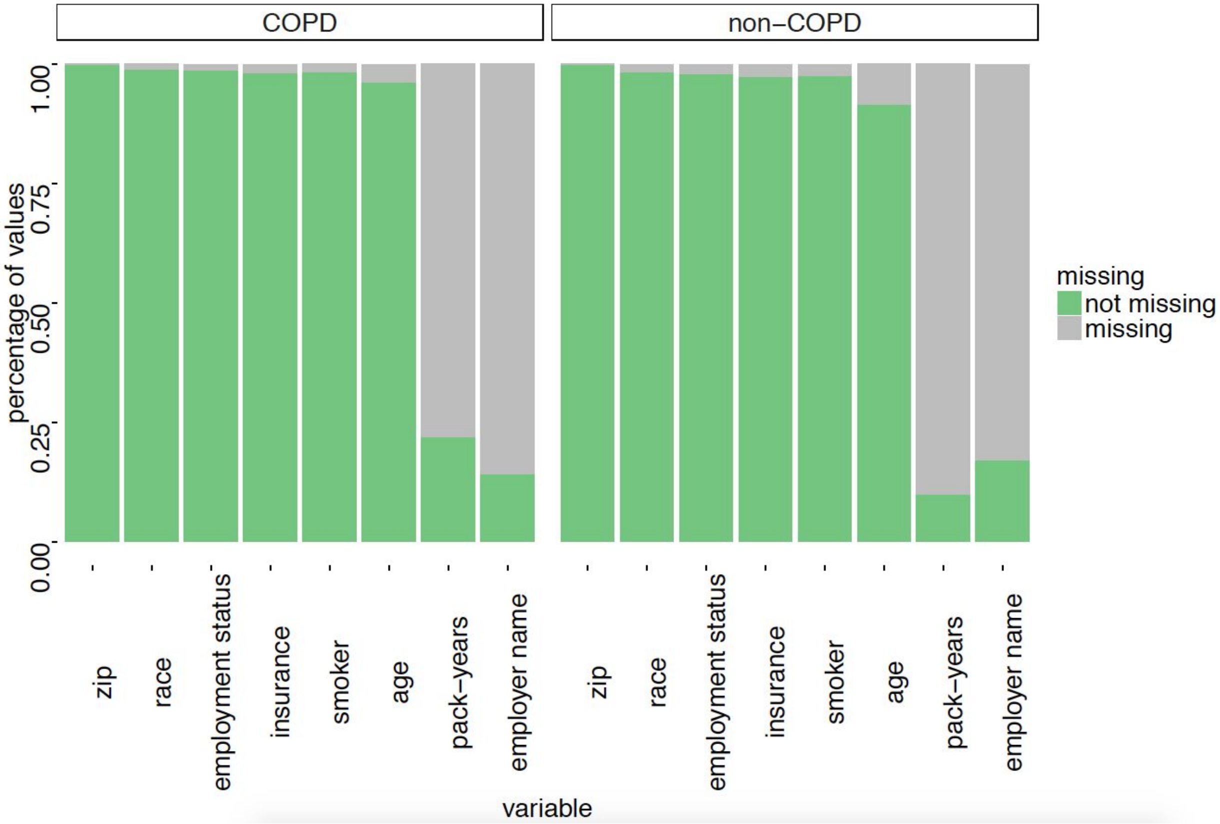
## Viewing Data Network FAA_DAILY

| ID | Name | County | State | Lat | Lon | Elevation (feet) | Start | End |
|---|---|---|---|---|---|---|---|---|
| *KABE* | ALLENTOWN | LEHIGH | PA | 40.650 | -75.440 | 376.0 | 1948-02-01 | 2016-04-07 |
| *KAOO* | ALTOONA | BLAIR | PA | 40.290 | -78.320 | 1504.0 | 1977-01-28 | 2016-04-07 |
| *KBVI* | BEAVER FALLS | BEAVER | PA | 40.770 | -80.390 | 1230.0 | 1996-01-02 | 2016-04-06 |
| *KBFD* | BRADFORD | MCKEAN | PA | 41.800 | -78.640 | 2142.0 | 1957-07-01 | 2016-04-07 |
| *KBTP* | BUTLER | BUTLER | PA | 40.770 | -79.950 | 1250.0 | 1992-02-26 | 2016-04-07 |
| *KCXY* | CAPITAL CITY | YORK | PA | 40.220 | -76.850 | 340.0 | 0000-00-00 | 2016-04-07 |
| *KFIG* | CLEARFIELD | CLEARFIELD | PA | 41.040 | -78.410 | 1516.0 | 2000-12-31 | 2016-04-07 |
| *KDYL* | DOYLESTOWN | BUCKS | PA | 40.330 | -75.120 | 394.0 | 1999-07-28 | 2016-04-07 |
| *KDUJ* | DUBOIS | JEFFERSON | PA | 41.180 | -78.900 | 1814.0 | 1973-01-27 | 2016-04-07 |
| *KERI* | ERIE | ERIE | PA | 42.080 | -80.170 | 730.0 | 1926-01-01 | 2016-04-07 |

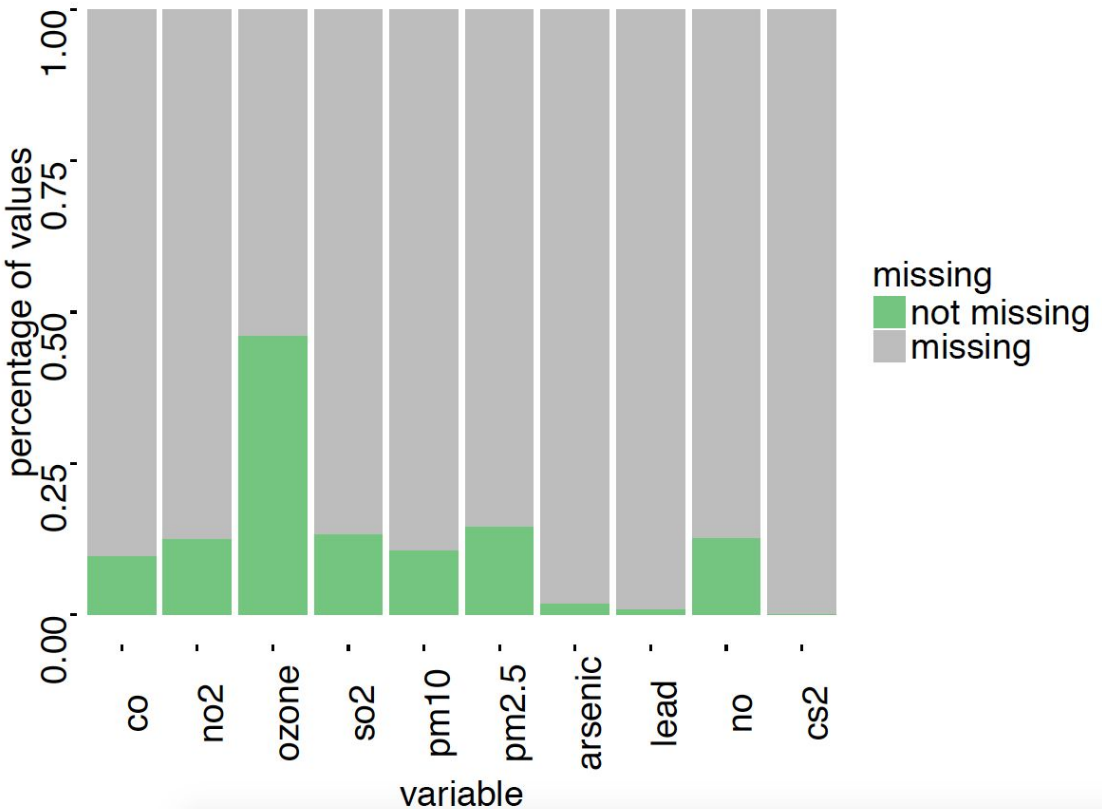
Sounds great!

So what’s the problem?

# The smoking and employment information was missing!



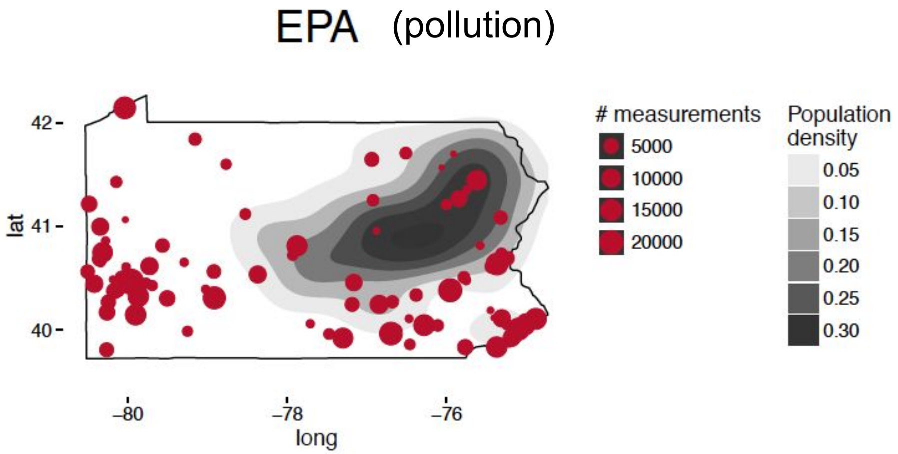
# The EPA “daily” values were not daily at all…

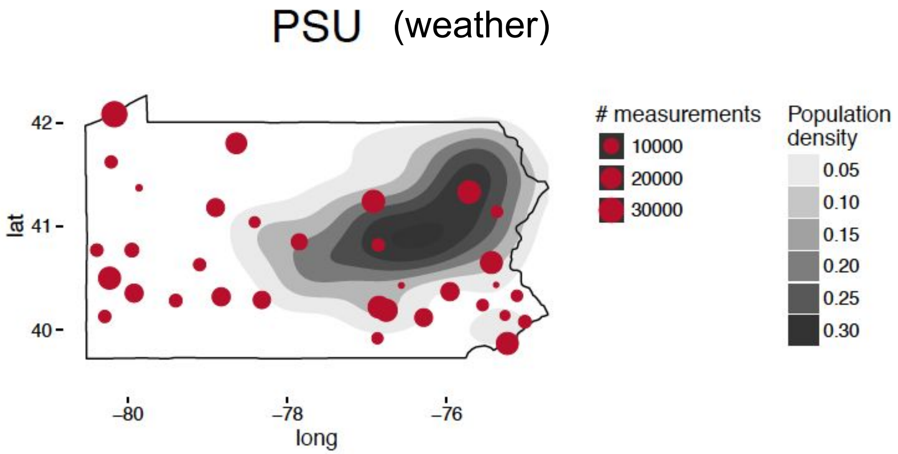


We went ahead and blended anyway…

# Data Blending

- Blended EPA and PSU with Geisinger using **date** and **closest zipcode.**
- **Problem:** not a lot of geographical overlap between Geisinger patients and environmental measurement sites.



# Pre-processing the data before modeling:

# Dealing with missing values

# How can we deal with missing values?

## Possible ideas:

1. **Exclude all** observations that had any **missing features** to only leave a modeling dataset with no missing data.
2. Utilize methods that **directly allow for missing data** in the modeling process.
3. **Exclude** all features that have more than a **threshold proportion** of missing values.
4. Perform **imputation** on all missing features using the **median**, mean or a k-nearest-neighbors approach from non-missing values from the same feature.

# How did we deal with missing values?

**Our approach:** to minimise data loss and ensure practicality

- **Remove** all variables with more than 8% missing values.
- **Impute** the remaining missing values
    - Numerical features: impute using the median.
    - Categorical features: impute using the mode.

# Pre-processing the data before modeling:

# Dealing with unbalanced classes

# How can we deal unbalanced classes?

- Many machine learning algorithms are known to perform poorly under class imbalance.
  - We have 5,704 COPD patients and 2,374 non-COPD patients.

## Possible ideas:

- **Upsample:** randomly sample labels from the smaller class (non-COPD) with replacement to be equal in number to the non-COPD labels.
- **Downsample:** randomly sample labels from the larger class (COPD) to be equal in number to the non-COPD labels.

# How did we deal unbalanced classes?

## Our approach

- **Upsample:** randomly sample labels from the smaller class (non-COPD) with replacement to be equal in number to the non-COPD labels.

No need to sacrifice sample size.

# Stepwise Feature Inclusion

# Stepwise Feature Inclusion

- **Geisinger Clinical**
  - Gender, marital status, employment status, age, race, asthma
- **Geisinger Clinical + Smoking**
  - Gender, marital status, employment status, age, race, asthma
  - *binary smoking variable*
- **Geisinger Clinical + Smoking + PSU weather data**
  - Gender, marital status, employment status, age, race, asthma
  - binary smoking variable
  - *average temperature, pressure and humidity in the week preceding the admission*

# Modeling

# We used empirically well-tested non-parametric models

- Random Forest
- GBM
- XGBoost

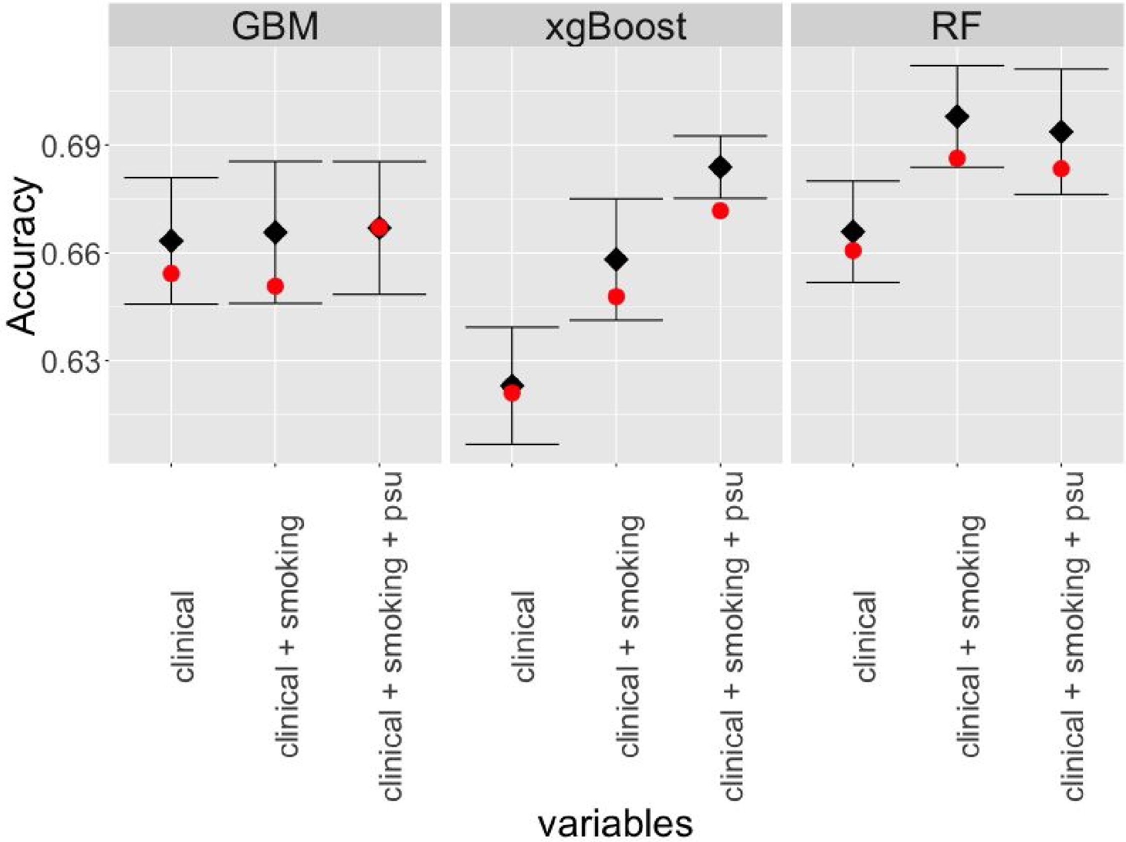
## To fit these models we used the R **caret** package

- Test interaction of various combination of input parameters e.g. for GBM varied interaction depth and number of trees
- Parameters selected using 5 repeated rounds of 10-fold CV

# Results

# Results

- **Black diamond:** average of CV estimates for the optimal parameter set.
- **Red circle:** prediction accuracy on withheld test set.




**Best performing model:**

- Random Forest
- **Accuracy** of **70%**

# Conclusion

# Conclusion

**Question:** For a new patient who has been diagnosed with pneumonia, do they have COPD?

- Data collected was plagued by missing values.
- Better performance accuracy may have been achievable with better quality data:
  - complete **smoking pack- years**
  - outdoor and indoor **pollution data**