Integrated Data Analysis for Early Warning of Lung Failure

Geisinger Health Collider Project: Stage 2

The Outliers: Rebecca Barter and Shamindra Shrotriya
Department of Statistics, UC Berkeley

June 28, 2016

Abstract

Chronic obstructive pulmonary disease (COPD) is a major cause of mortality worldwide, with approximately 12 million adults in the U.S. having been diagnosed with COPD. Our aim is to develop methods capable of effectively predicting cases of undiagnosed COPD among those whose primary reason for hospitalization was pneumonia. Most existing algorithmic approaches to similar prediction problems focus only on utilizing clinical information. Our approach, however, aims to incorporate external environmental data sources that are not captured by the clinical records using a process called "data blending". We also investigate several leading supervised machine learning algorithms including Random Forest, Gradient Boosting Machines (GBM) and eXtreme Gradient Boosting (XGBoost) to improve COPD classification accuracy. We find that smoking and weather information significantly improve the predictive power of these algorithms in terms of predicting COPD among pneumonia patients.

Keywords. COPD, pneumonia, random forest, Gradient Boosting Machines (GBM), eXtreme Gradient Boosting (XGBoost)

1 Introduction

Chronic obstructive pulmonary disease (COPD) is a major cause of mortality worldwide, with approximately 12 million adults in the U.S. having been diagnosed with COPD (Lozano et al., 2012). Crucially, however, it is estimated that a further 12 million adults in the U.S. are currently living with undiagnosed COPD (NIH, 2010). Individuals with undiagnosed COPD typically experience related adverse health effects, and one of the primary reasons for hospitalization amongst undiagnosed COPD patients is pneumonia. Among patients with a secondary diagnosis of acute exacerbation of COPD, pneumonia was the primary reason for hospitalization for 22.3 percent of admissions (Wier et al., 2011). Our aim is to develop methods capable of effectively predicting cases of undiagnosed COPD amongst those whose primary reason for hospitalization was pneumonia. Most existing algorithmic approaches to similar prediction problems focus only on utilizing clinical information. Our approach, however, aims to incorporate external environmental data sources that are not captured by the clinical records using a process called "data blending".

2 Risk factors for COPD

According to the Mayo Clinic website, the most prominent risk factors for COPD are exposure to tobacco smoke, smoking combined with asthma, occupational exposure to dust and chemicals, age and genetics. Our clinical data provides us with information on tobacco smoke, asthma and age,

however we do not have access to data on dust and chemical exposure or genetic information. For this project, our goal was to explore the predictive power of information of smoking (sourced from Geisinger patient records), outdoor pollution and weather (sourced from environmental agencies), and occupational exposure to dust and chemicals (inferred by employment information sourced from Geisinger patient records) on the prediction of COPD among patients who have been diagnosed with pneumonia.

2.1 Smoking and age

Smoking is well known to be one of the major risk factors for COPD, however only 25% of smokers actually develop COPD (Lokke et al., 2006). Figure 1 displays two histograms from Geisinger's clinical data (described in the next section) of pack-years (the number of cigarette packs smoked daily multiplied by the number of years smoked) and for age. The red histograms correspond to the COPD population, whereas the blue histograms correspond to the non-COPD population. Notice that the distribution of pack-years for the COPD population has slightly fatter tails and is shifted further to the right than the distribution for the non-COPD population, indicating that pneumonia patients with COPD tend to smoke more than those without COPD. We will discuss below that a lot of the pack-years smoking data is missing (in fact, the county for which we have the largest proportion of smoking data is Sullivan, and even in this case we have only 65% of smoking data), however, we have almost complete binary smoker status information, so this figure presents a possibly biased view of the data. Further, in the second panel, we can see that there are few individuals younger than 40 who are diagnosed with COPD, whereas there are a decent number of non-COPD pneumonia patients in this age group.

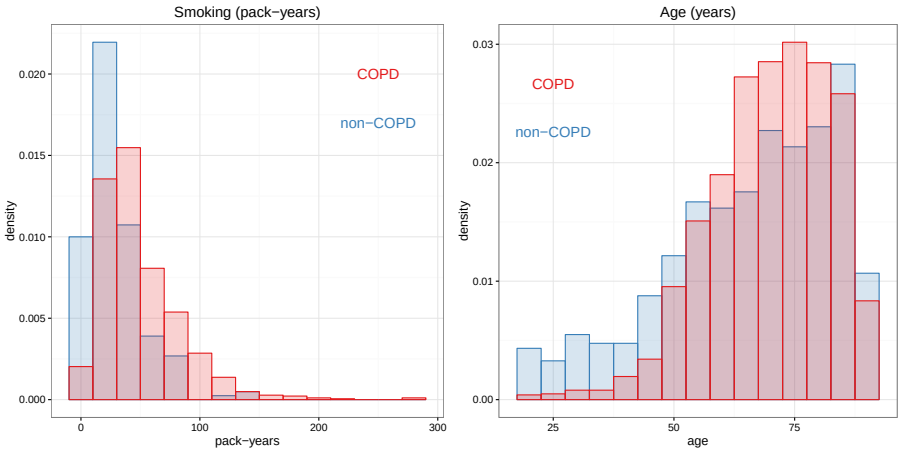

Figure 1: A comparison of the distribution of (1) pack-years, and (2) age for individuals in the COPD pneumonia subgroup (red) and individuals in the non-COPD pneumonia population (blue)

Next, Figure 2 displays two maps of Pennsylvania (restricting to the counties for whom we have at least 20 Geisinger patients in our filtered subset). The map on the left displays the proportion of individuals in each county who are smokers, and the map on the right displays the proportion of individuals in each county in our COPD sub-population. Although there are some counties which appear to have both higher proportions of smokers as well as higher COPD rates, this relationship certainly does not hold for all counties, implying that there is more contributing to the development of COPD than just smoking status.

2.2 Occupational exposure to VOCs

Pollutants such as Volatile Organic Compounds (VOCs) emissions from biomass fuels are typically found in household and workplace products, such as paints, wood preservatives, aerosol sprays,

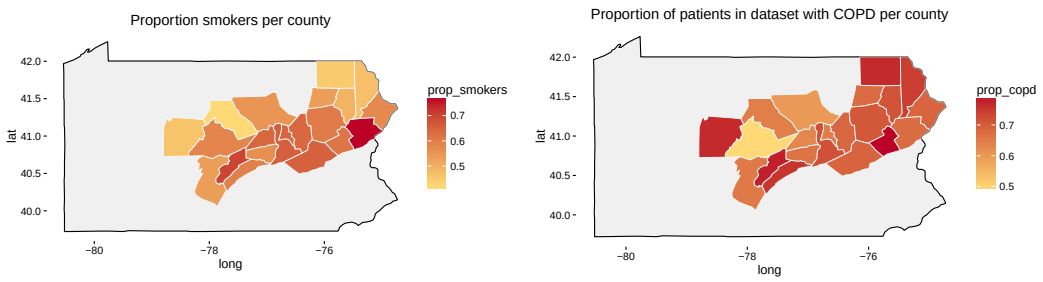

Figure 2: County maps (for the counties with at least 20 Geisinger Patients) of Pennsylvania displaying (1) the proportion of smokers per county, and (2) the proportion of individuals in the county that are in the filtered COPD sub-population

cleaners and disinfectants, etc, and have in previous studies been linked to the development of COPD (Van Berkel et al., 2010). However, studies examining the effect of VOCs on COPD were primarily examining the composition of VOCs in the exhaled breath of COPD patients, rather than in the air.

Since, we do not have access to the VOC composition of the exhaled breath of our patients, we hoped to use employment information to infer whether individuals worked in an environment which would expose them to high levels of VOCs. Unfortunately, as we will discuss below, the employment information available only covered approximately 10% of patients, rendering us unable to explore this avenue of investigation.

2.3 Outdoor pollution

Aside from the occupational exposure to dust and chemicals, which must be measured in the workplace or in the home, there exists limited evidence of other types of environmental risk factors such as air pollution (Ko and Hui, 2012). Thus, since we were unable to examine individual exposure to VOCs in the home or workplace, our next course of investigation was to obtain outdoor VOC (and other pollutant) levels by zip-code. The hypothesis is that individuals who live in an area with higher levels of pollutants in the air were more likely to develop COPD.

Although we were able to find "daily" state-wide Environmental Protection Agency (EPA) data, this data was very far from being at the zip-code, or even county, level, and, as we will discuss below, had so many missing values that it became unsalvageable.

2.4 Weather

Our final hypothesis is that individuals who are diagnosed with pneumonia in warmer weather are more likely to have COPD. In particular, there are more pneumonia diagnoses that occur in colder weather (primarily due to individuals spending more time in enclosed areas, and thus having more exposure to pneumonia), so that a pneumonia diagnosis that occurs in colder weather is less likely to be due to COPD.

Figure 3 displays a plot of the temperature overlaid with the pneumonia admissions per month based on data collected from Pennsylvania State University (PSU); see below. It is clear that there is a seasonal trend wherein in lower temperatures, there are more pneumonia admissions than in higher temperatures. Ideally, we would have liked to obtain zip-code level weather data, however, the PSU data suffers from the same problem as the EPA data in that the data collection sites are not located in regions densely populated by Geisinger patients (this point will be discussed later). However, unlike the EPA data, the PSU data does not suffer from severe issues of missing values.

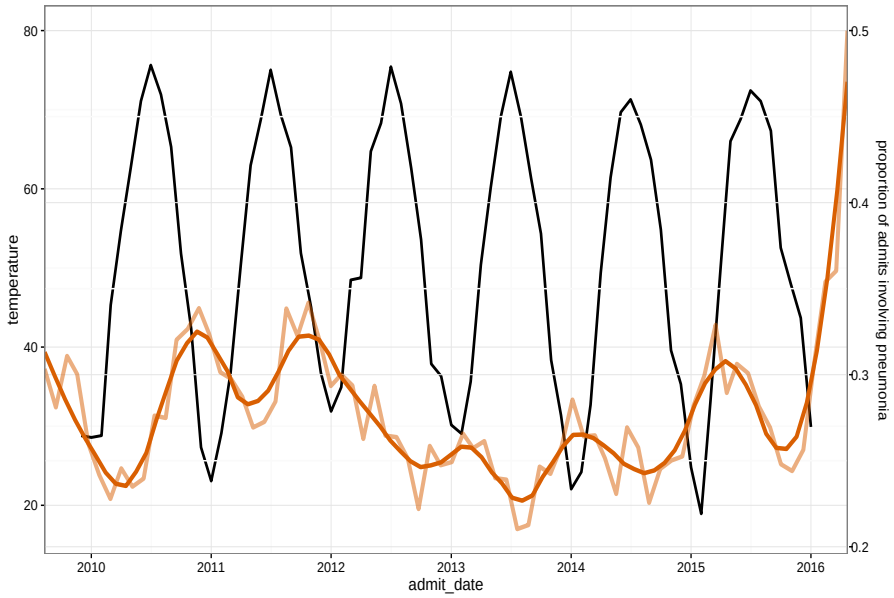

Figure 3: Pneumonia admissions for all patients by month (raw jagged orange line and smoothed orange trend line) as well as the average monthly temperature (black dotted line) obtained from the PSU Climate data.

3 The clinical data

The clinical data was collected from medical records on patients of the Geisinger Health System in Pennsylvania. The dataset contains longitudinal data on patient visit information over approximately 6 years from December 2009 to January 2016 for a total of 31,108 individuals (who have either been diagnosed with COPD or have been diagnosed with pneumonia), among whom, 19,721 have a discharge diagnosis of COPD. The clinical dataset contains 46 unique variables, 24 of which contain diagnosis information, and the remainder of which contain fixed (non-longitudinal) demographic information, such as gender, age, race, marital status, zip code, religion, language, and employer information. In addition to the clinical data supplied by Geisinger, we also have information on smoking which we have combined into a binary *smoker* variable (equal to 0 if the individual has never been recorded as a cigarette smoker, and equal to 1 otherwise) as well as *pack-years* variable, defined by the average number of packs smoked per day multiplied by the number of years for which the individual smoked.

3.1 Filtering the clinical data

Our goal is to develop a classification algorithm capable of predicting, for a new patient who has been diagnosed with pneumonia, whether or not they have COPD. We thus filter the data so that our two classes (COPD and non-COPD) encompass only those who have been diagnosed with pneumonia. Further, the majority of the clinical information is not longitudinally interesting (variables such as gender, zip-code, etc), and thus in order to avoid issues arising from multiple observations from individual patients (where, in most cases, the only differences between observations taken at distinct visits are in the date-related and diagnosis variables), we filter the data to contain only data from a single visit per patient.
The exact filtering process is outlined in Figure 4, and involves removing those aged below 18 and above 90 (note that the ages of patients aged 90 and above are recorded in the data simply as > 90), and those who visited Geisinger but live outside of the state of Pennsylvania. Next, since we

are interested in predicting undiagnosed COPD for patients *who are diagnosed with pneumonia*, we remove the COPD patients who have not been diagnosed with pneumonia in the 6-year period. We also remove the non-COPD patients (all of whom have been diagnosed with pneumonia at some point), who have been diagnosed with pneumonia more than once in the 6-year period (since these people may have undiagnosed COPD).

At the end of this filtering process, we filter the data so that each patient provides a single observation. To do this, we first filter the data remove all visits at which there was no pneumonia diagnosis. If there was only one pneumonia visit, then we take the data from this visit as our record for the patient. If, however, there were multiple such visits, we take the visit at which the individual first presented with pneumonia.

Finally, Prior to conducting exploratory and predictive analysis, we randomly split our patients into three subsets: a training set (70% of the original data; a total of 24,4887 patients, 9,110 without COPD and 15,777 with COPD) and a testing set (the remaining 20% of the original data; a total of 6,221 patients). All exploratory data analysis (EDA) is based on the patients in our training dataset, while the predictive accuracy is based on models built using the patients in the training set but is tested using patients in the test set.

Following the filtering steps described above, we find that in our *training set* there are $4,765$ individuals in our COPD population and $2,067$ individuals in our non-COPD population.

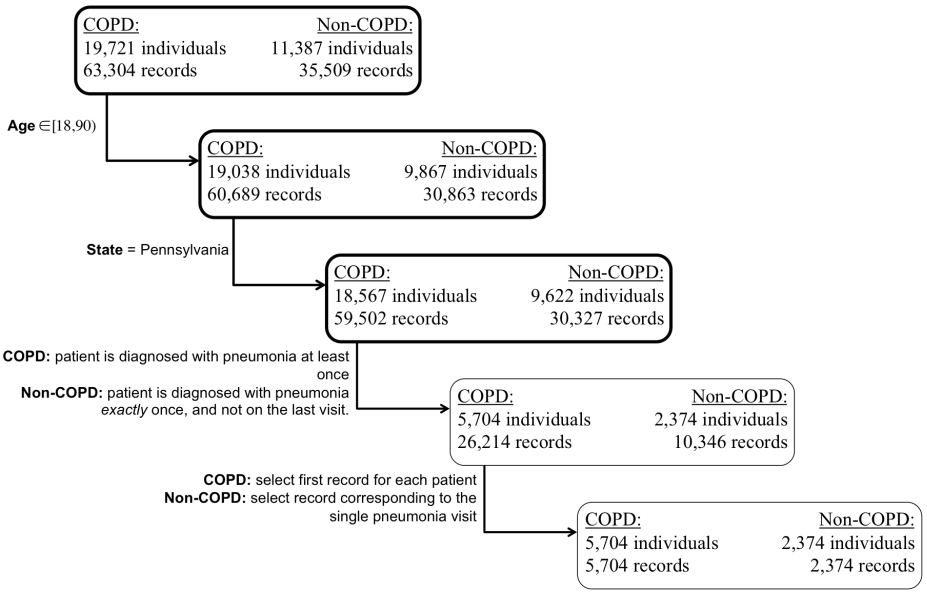

Figure 4: A flow chart describing the filtering process for the Geisinger clinical data. The filtering step is presented to the left of the arrow, and the number of patients and corresponding number of total records in the dataset are reported for each filtering step. The size of the data is represented by the thickness of the box around the description

3.2 Missing values

Unfortunately the clinical data suffers from severe missingness for several of our most informative variables (Figure 5). For patients aged 18 and over, smoking pack-years data is available for only 11% of the filtered non-COPD patients and 20% of the COPD patients. As smoking is one of our primary COPD predictors, this issue is particularly irksome, however fortunately, we do have a binary smoking variable for over 99% of patients (this variable is defined to be 0 if the patient has never been recorded as a cigarette smoker over the 6 year period, and is defined to be 1 otherwise). Further, our goal of inferring biomass fuel exposure from employment information is significantly hindered by the fact that we have employment information for only 15% of non-COPD and 10%

of COPD patients aged 18 and over.

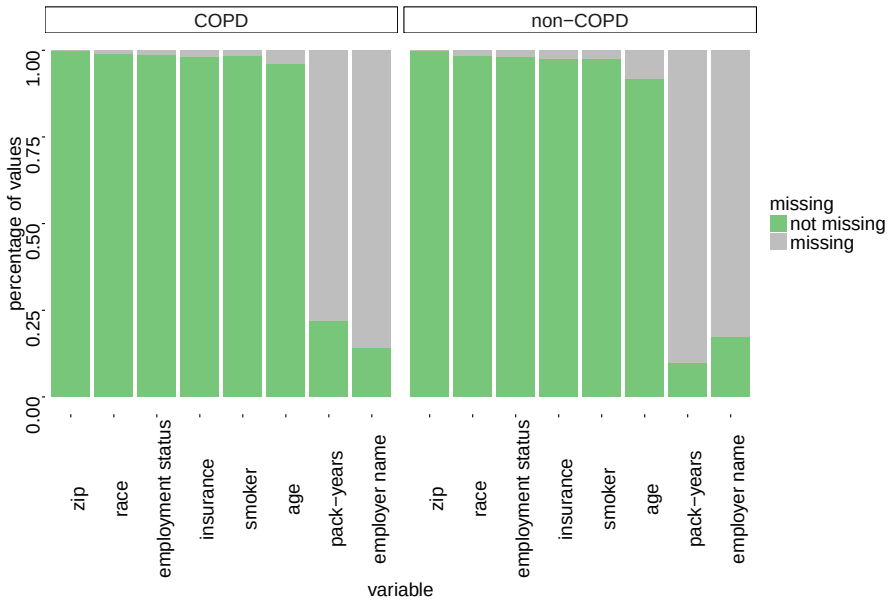

Figure 5: The proportion of measurements made that are missing for the clinical variables.

4 The environmental data

4.1 EPA data

The first source of environmental data was collected from the United States Environmental Protection Agency (EPA) and contains "daily" pollutant and climate measurements in the following four categories:

- **Criteria pollutants**: ozone, SO_2, CO, and NO_2. These pollutants are known to cause smog, acid rain, and other health hazards,
- **Inhalable particulates**: PM2.5 and PM10; particles which pass through a size-selective inlet with a 50% efficiency cut-off at 2.5 and 10 μm aerodynamic diameter, respectively
- **Toxics, Precursors, and Lead**: hazardous air pollutants (HAPs), volatile organic compounds (VOCs), nitrous oxides (NONOxNOy), and lead.
- **Meteorological measurements**: resultant winds, temperature, barometric pressure, and relative humidity.

As the EPA daily data measurements appeared to be a very promising source of information for both air pollution and weather data over both time and various geographical location, we spent several weeks obtaining, cleaning, and eventually blending this data to our clinical data from Geisinger (see below for the blending method). However, having obtained our final blended version of the data, we began to discover that the daily EPA data was not collected daily at all, rather, there are patterns of extreme missingness, and the geographical measurement locations were far from those of the bulk of the Geisinger patients, as we will discuss below.

4.2 Pennsylvania State University Climatologist data

Although the EPA data provided us with meteorological measurements, as we will see below, these measurements are collected sporadically, and for some variables, there is no data prior to 2011. As an alternative source of meteorological measurements, we collected data from the Pennsylvania State University (PSU) Climatologist website, which contained data for temperature, dew point temperature, precipitation (inches), wind speed, wind direction and humidity.
Although the PSU data did not suffer from the same problems of missing values as the EPA data, we similarly found that the distribution of the data collection sites were far from the locations of the Geisinger patients.

4.3 Blending the environmental data and the clinical data

Each measurement contained in the EPA and PSU dataset is associated with a date and a latitude/longitude location. The Geisinger data, on the other hand, provided us only with patient ZIP codes, not latitude and longitude information. Although converting from latitude and longitude to zip code is fairly straightforward, few of the Geisinger patient zip codes are actually represented in the EPA and PSU datasets. Thus, our goal was: for every Geisinger patient, find the closest EPA site using the Vincenty (sphere) distance metric on the (latitude, longitude) measurements, and take the environmental measurements from that site for each date (if an EPA or PSU measurement was not recorded for a specific date, the measurement was reported as missing).
Figure 6 displays the distribution of Geisinger patients (the 2D density is displayed as a grey contour map) and the corresponding distribution of EPA and PSU measurement sites. Unfortunately, we found that the bulk of the environmental measurements were taken in locations far from where the majority of the Geisinger patients live. As a result, the blending process results in 65% of all patient's EPA measurements being sourced from only 5 sites.

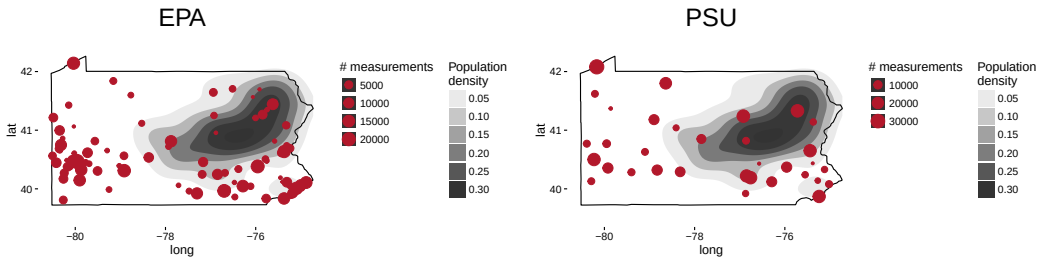

Figure 6: A map of Pennsylvania displaying the EPA data collection locations as well as the distribution of geisinger patients.

4.4 Missing EPA data

As we have hinted in earlier sections, many of our efforts were hindered by significant amounts of missing values in the EPA data. Figure 7 presents the number of measurements taken for the environmental variables per month. What we find is that for the criteria gases, prior to 2011, ozone measurements were only taken in the months of March-October. Further, since 2014 the number of CO measurements have dropped below 20 per month per location. Next, for the particulates measurements, both PM10 and PM2.5 measured with the Federal Reference Method (FRM) sampler are recorded an average of 20 times per month per location (while measurements for PM2.5 taken using non-FRM methods are recorded on average less than 10 times per month per location).
For the meteorological measurements, pressure is not measured at all prior to 2011, after which it is recorded almost every day per location. Similarly, prior to 2011, there were only 3-4 tem-

perature measurements made per month and between 2011 and 2014 there are 9-10 temperature measurements made per month, after which the number of measurements per month increased to above 20. Wind speed is measured sporadically from mid-2009 until the beginning of 2010. Between 2010 and 2011, wind is not measured at all, and is subsequently measured approximately 8 times per month after 2011. Finally, the toxics, precursors, and lead measurements are taken only a few times per month, with the exception of NO, which is measured approximately 20 times per month.

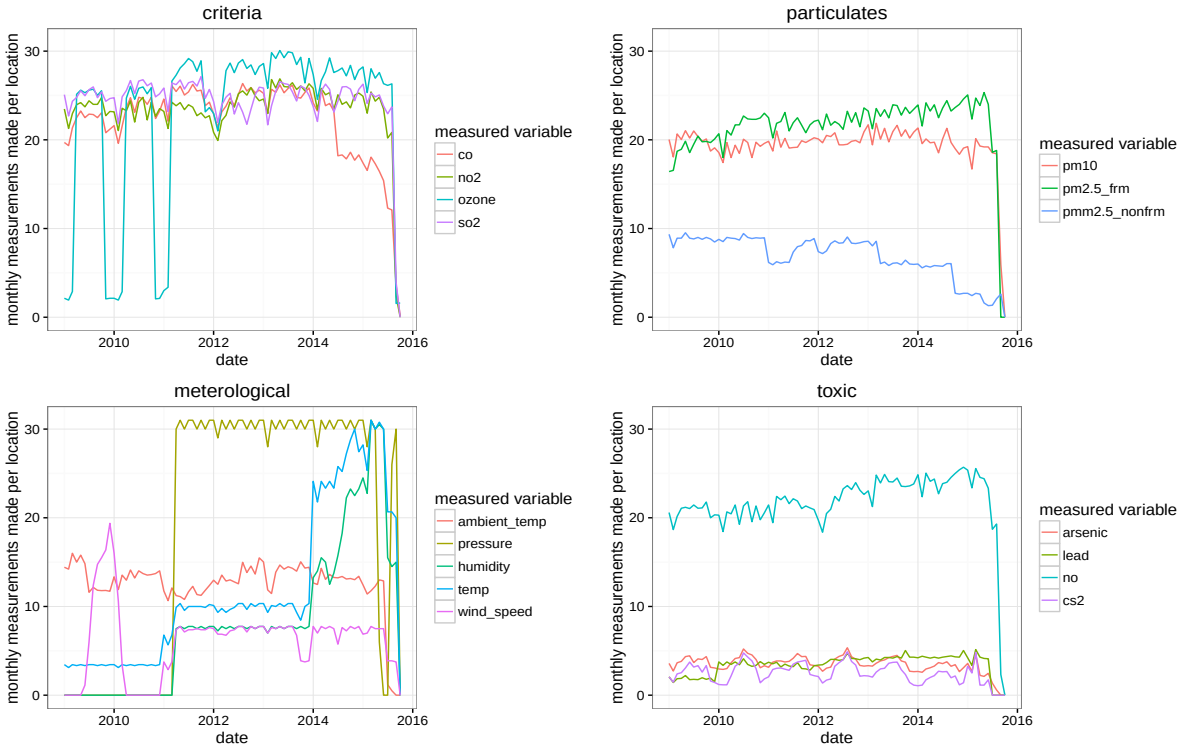

Figure 7: The number of measurements made for each environmental variable per month divided by the number of sites (locations) where the measurements are taken. The thickness corresponds to the number of different locations at which measurements are made for that variable.

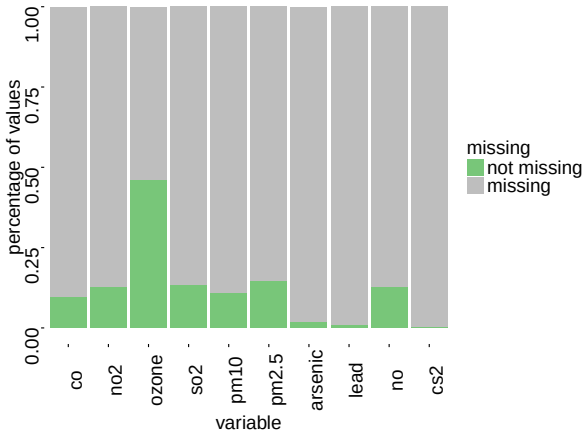

Figure 8: The proportion of measurements made that are missing for the EPA variables.

Fortunately we have daily climate measurements from each site for all of our PSU variables with the exception of precipitation and maximum wind gust, for which we have on average of 15 and 18 measurements per month respectively.

5 Imputation

As discussed, our clinical data had several features which later combined with EPA and PSU had varying degrees of missing data. This presented several issues for accurate modeling purposes. To correct for this we had the following four broad approaches to deal directly with the missing data:

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

Given these broad approaches and noting that our combined training data had less than 7000 observations we decided to avoid approach 1 altogether in order to retain as much modeling data as possible.
Also, for modeling purposes we intended to use non-parametric algorithms supervised learning algorithms including Random Forests (RF), Generalized Boosting Machines (GBM) and eXtreme Gradient Boosting (XGBoost). In general we observed that the `R` implementations of these machine learning algorithms did not allow us to easily deal with missing data with a fine degree of control. This ruled out using approach 2 alone for dealing missing data.
As such, we ended up using a combination of approaches 3 and 4 above to impute the missing data in our combined training and validation datasets. The specific process is as follows:

1. Exclude all features which had more than 8% of their values missing. The 8% threshold was determined using trial-and-error to ensure that key internal and external data features remained in the dataset for modeling after this exclusion criteria.
2. For the remaining features, we imputed using the following simple rules:
    - **numerical features**: impute missing values using the median of all non-missing values for these individual features.
    - **categorical features**: impute missing values using the mode (most common value) of all non-missing values for these individual features.

After imputation we found that for all of the remaining features after imputation that there was a mere 0.03% increase in the mean value compared to the same features before imputation. This gave us confidence that our imputation process (as described) had not caused any major instability issues in the modeling data.
We also believe that the imputation approach is highly practical in a Geisinger hospital setting where some clinical data features may not be collected (i.e. missing) for some patients. By imputing this data we ensure that this is not a a barrier for modeling and predictive purposes.

6 Modeling

After data cleaning and imputation, our fundamental goal was to build a high accuracy binary classifier for COPD using features from clinical Geisinger data and external blended weather data. The following sections detail our models fundamental approach to the classification process including models used, train-validation methodology, approach to testing external data in a stepwise manner, model parameters varied and key results.

6.1 Models

In order to get highest accuracy (among other metrics) in our binary classification of COPD we opted to use the following well tested supervised learning models in our classification process:

1. **Random Forests (RF)**: an ensemble method based on trees. The idea is to choose randomized samples of the features and build a number of tree algorithms on this subset of features. The predicted label for a tree is chosen according to a majority vote in the corresponding leaf of the tree. Since trees have low bias, taking the average of the bootstrapped trees reduces the variance, assuming that the features are not correlated. By randomizing the choice of features the trees are almost i.i.d and thus the algorithm can be seen as an improvement of Bagging.

2. **Gradient Boosting Models (GBM)**: an ensemble method of weak learners (decision trees) which typically have high bias and low variance. The boosting process aims to reduces error mainly by reducing overall bias (and also the variance) by aggregating the output from many fitted trees.

3. **eXtreme Gradient Boosting (XGBoost)**: can be viewed as an extension of GBMs which use a more regularized model formalization to better control for over-fitting.

It is important to note that better data quality has been empirically shown to give better incremental modeling accuracy than just using more sophisticated machine learning algorithms. Nonetheless we opted to use a variety of well-tested supervised learning algorithms to ensure highest accuracy after ensuring a clean modeling dataset with the richest clinical and blended features.
For all models we randomly split our processed and imputed dataset into training and validation sets. In order to avoid overfitting on the combined training dataset we used 10-fold cross validation and repeated the process 5 times for each model. We then chose the best model of each type to run against the validation dataset to determine the overall model of best fit based primarily on COPD classification accuracy.

6.2 Class imbalance of labels

In our training set, we had a slight class imbalance in favor of COPD labels compared to non-COPD labels. In order to address the imbalance issue in the training process we investigated 2 approaches:

1. **Downsampling** the observations in the combined training set so as to randomly sample labels from COPD to be equal in number to the non-COPD labels.

2. **Upsampling** the observations in the combined training set so as to randomly oversample (bootstrap) labels from non-COPD to be equal in number to the COPD labels.

After running a few tests, empirical evidence showed that upsampling provided slightly higher accuracy compared to downsampling (and avoided the need to reduce the training dataset size). The RF, GBM and XGBoost `R` implementations all allowed upsampling to occur within the tree building process as an option rather than as a pre-processing requirement on the training dataset.

6.3 Stepwise feature selection

One of the key objectives was to investigate the impact and influence of the external data in the classification process. In order to achieve this we decided to include features in a stepwise manner in the modeling process for the RF, GBM and XGBoost models as follows:

1. **Clinical**: Run classifiers using Geisinger clinical data only (gender, marital status, employment status, age, race, asthma)

2. **Clinical + smoking**: Run classifiers using Geisinger clinical data and a binary smoking variable (sourced from Geisinger patient records)

3. **Clinical + smoking + PSU**: Run classifiers using Geisinger clinical data, smoking data and PSU weather data (average temperature, pressure and humidity in the week preceding the admission)

By approaching the problem in this stepwise manner we could try to best isolate the effects of the additional features in a more controlled manner.

6.4 Model Parameter Interactions

We ran all three model types (RF, GBM and XGBoost) using broad parameter sets in order to identify to optimal parameter values for each model based on cross-validated performance measures. The key parameter values in each model are summarized below:

GBM parameters	Description	Values Tested
interaction.depth	number of splits it has to perform on a tree	1, 5, 9
n.trees	number of trees	30, 60, 90, 120, ..., 570, 600
shrinkage	a shrinkage parameter applied to each tree	0.1
n.minobsinnode	min num of observations in terminal nodes	3

Table 1: Parameters for GBM

RF parameters	Description	Values Tested
mtry	number of predictors sampled for spliting at each node	1, 2, 3, ..., 15

Table 2: Parameters for RF

XGBoost parameters	Description	Values Tested
nrounds	max number of iterations	1000
eta	step size of each boosting step	0.05, 0.1
max_depth	maximum depth of the tree	2, 8, 14
gamma	model complexity parameter	0
colsample_bytree	subsampling parameter	1.0, 0.5
min_child_weight	model complexity parameter	1

Table 3: Parameters for XGBoost

All models were fit using an interaction of all parameter values listed above during the training process and ultimately the model with the highest classification accuracy was selected from RF, GBM and XGBoost types and run on the validation dataset.

7 Results

Having run the models described above, we found (Figure 9) that for the XGBoost and RF algorithms, the addition of the binary smoking variable increased the prediction accuracy by 2.5% when evaluating the performance on the withheld test set. Adding the PSU environmental data further increased the test set accuracy by another 2.5% and decreases the variance for the XGBoost algorithm. For the RF algorithm, however, the accuracy is very slightly decreased by the addition of the weather variable.
The GBM model, on the other hand, performs similarly in the cross-validation accuracy for all three variable sets, however, the GBM model performs the best on the validation set when including

the smoking and PSU weather data. The most accurate algorithm is RF with clinical and smoking data which obtained an average CV accuracy of 70% on the training set and an accuracy of 68% on the withheld test set.

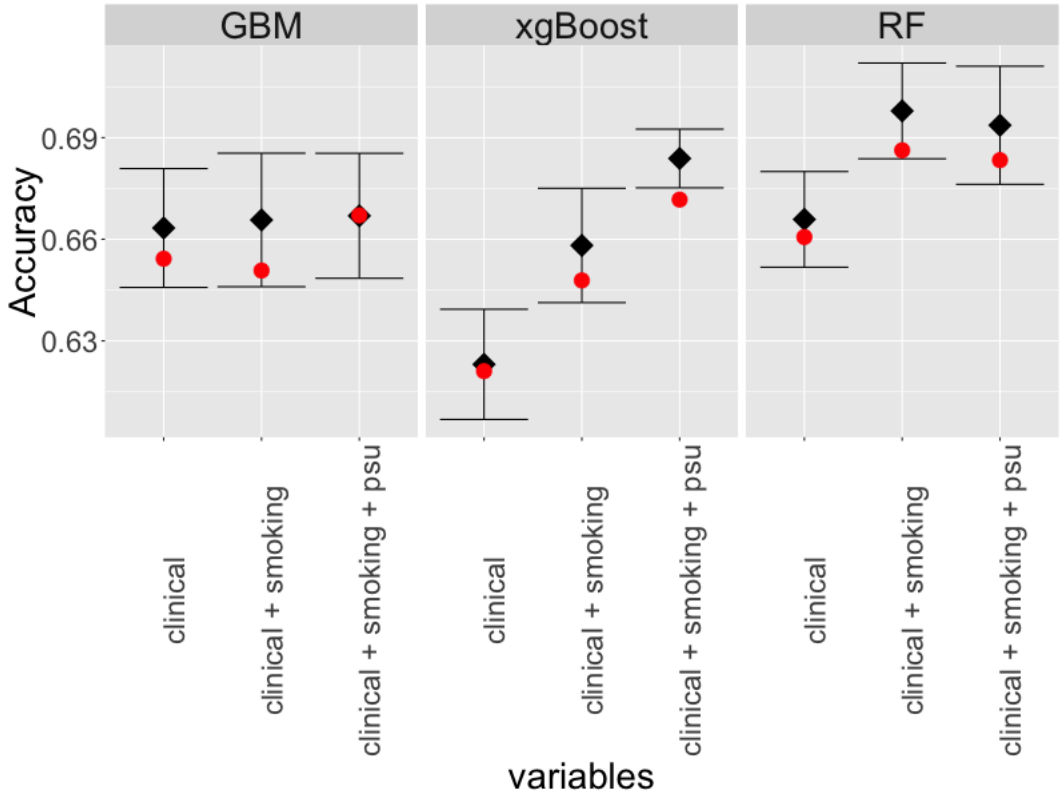

Figure 9: The black diamonds display the cross-validation point estimate of classification accuracy based on the average of 5 repeated runs of 10-fold cross-validation on the training data for the model fit with the optimally tuned parameters (identified by the parameter set that yielded the highest CV accuracy). The error bars correspond to ± 1 SD based on the CV accuracy estimates. The red circle corresponds to the performance of the model on a withheld test set.

We hypothesize that had we been able to collect better quality data (for example complete pack-years and outdoor and indoor pollution data), we would have obtained more accurate predictive results.

8 Conclusion

Our goal was to develop algorithms that, for a patient who had been diagnosed with pneumonia, could combine clinical and environmental data in order to predict whether the patient has COPD or not. We obtained clinical and smoking data from Geisinger and environmental data from the Environmental Protection Agency and Pennsylvania State University. Unfortunately, we found that there was limited smoking pack-years data and employment data present in the Geisinger records, and the EPA data also exhibited high levels of missingness. The PSU weather data was fairly comprehensive, but suffered somewhat from a lack of geographical overlap with the Geisinger population.

Following imputation, we built three algorithms for each stepwise combination of clinical, smoking and PSU weather data, and found that in for all three algorithms, the addition of smoking and weather variables improved training set performance compared with the algorithms without these

variables. The performance of these algorithms on a withheld test set was also improved upon the incorporation of weather data when compared with the algorithm that did not include the weather data in all cases. Unfortunately, however, despite these improvements, the best accuracy rate achieved was 70%. Future work would examine obtaining higher quality features for example from complete environmental pollution data and smoking pack-years information, with the goal of improving these accuracy rates even further.

9 Acknowledgments

We would like to thank Dr. Nicholas Marko, Dr. Joseph Klobusicky, Dr. Oleg Roderick, Dr. Jason Brown and Mr. Debdipto Misra of the Marko Lab for their kind assistance with providing Geisinger data and clarifying our numerous related queries. Their useful comments helped greatly improve our analysis and approach.

References

Ko, F. W. S. and D. S. C. Hui (2012, April). Air pollution and chronic obstructive pulmonary disease. *Respirology (Carlton, Vic.) 17*(3), 395–401.

Lokke, A., P. Lange, H. Scharling, P. Fabricius, and J. Vestbo (2006, November). Developing COPD: a 25 year follow up study of the general population. *Thorax 61*(11), 935–939.

Lozano, R., M. Naghavi, K. Foreman, S. Lim, K. Shibuya, V. Aboyans, J. Abraham, T. Adair, R. Aggarwal, S. Y. Ahn, M. Alvarado, H. R. Anderson, L. M. Anderson, K. G. Andrews, C. Atkinson, L. M. Baddour, S. Barker-Collo, D. H. Bartels, M. L. Bell, E. J. Benjamin, D. Bennett, K. Bhalla, B. Bikbov, A. Bin Abdulhak, G. Birbeck, F. Blyth, I. Bolliger, S. Boufous, C. Bucello, M. Burch, P. Burney, J. Carapetis, H. Chen, D. Chou, S. S. Chugh, L. E. Coffeng, S. D. Colan, S. Colquhoun, K. E. Colson, J. Condon, M. D. Connor, L. T. Cooper, M. Corriere, M. Cortinovis, K. C. de Vaccaro, W. Couser, B. C. Cowie, M. H. Criqui, M. Cross, K. C. Dabhadkar, N. Dahodwala, D. De Leo, L. Degenhardt, A. Delossantos, J. Denenberg, D. C. Des Jarlais, S. D. Dharmaratne, E. R. Dorsey, T. Driscoll, H. Duber, B. Ebel, P. J. Erwin, P. Espindola, M. Ezzati, V. Feigin, A. D. Flaxman, M. H. Forouzanfar, F. G. R. Fowkes, R. Franklin, M. Fransen, M. K. Freeman, S. E. Gabriel, E. Gakidou, F. Gaspari, R. F. Gillum, D. Gonzalez-Medina, Y. A. Halasa, D. Haring, J. E. Harrison, R. Havmoeller, R. J. Hay, B. Hoen, P. J. Hotez, D. Hoy, K. H. Jacobsen, S. L. James, R. Jasrasaria, S. Jayaraman, N. Johns, G. Karthikeyan, N. Kassebaum, A. Keren, J.-P. Khoo, L. M. Knowlton, O. Kobusingye, A. Koranteng, R. Krishnamurthi, M. Lipnick, S. E. Lipshultz, S. L. Ohno, J. Mabweijano, M. F. MacIntyre, L. Mallinger, L. March, G. B. Marks, R. Marks, A. Matsumori, R. Matzopoulos, B. M. Mayosi, J. H. McAnulty, M. M. McDermott, J. McGrath, G. A. Mensah, T. R. Merriman, C. Michaud, M. Miller, T. R. Miller, C. Mock, A. O. Mocumbi, A. A. Mokdad, A. Moran, K. Mulholland, M. N. Nair, L. Naldi, K. M. V. Narayan, K. Nasseri, P. Norman, M. O'Donnell, S. B. Omer, K. Ortblad, R. Osborne, D. Ozgediz, B. Pahari, J. D. Pandian, A. P. Rivero, R. P. Padilla, F. Perez-Ruiz, N. Perico, D. Phillips, K. Pierce, C. A. Pope, E. Porrini, F. Pourmalek, M. Raju, D. Ranganathan, J. T. Rehm, D. B. Rein, G. Remuzzi, F. P. Rivara, T. Roberts, F. R. De Len, L. C. Rosenfeld, L. Rushton, R. L. Sacco, J. A. Salomon, U. Sampson, E. Sanman, D. C. Schwebel, M. Segui-Gomez, D. S. Shepard, D. Singh, J. Singleton, K. Sliwa, E. Smith, A. Steer, J. A. Taylor, B. Thomas, I. M. Tleyjeh, J. A. Towbin, T. Truelsen, E. A. Undurraga, N. Venketasubramanian, L. Vijayakumar, T. Vos, G. R. Wagner, M. Wang, W. Wang, K. Watt, M. A. Weinstock, R. Weintraub, J. D. Wilkinson, A. D. Woolf, S. Wulf, P.-H. Yeh, P. Yip, A. Zabetian, Z.-J. Zheng, A. D. Lopez, C. J. L. Murray, M. A. AlMazroa, and Z. A. Memish (2012, December). Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet (London, England) 380*(9859), 2095–2128.

NIH (2010, October). Chronic Obstructive Pulmonary Disease. Technical report.

Van Berkel, J. J. B. N., J. W. Dallinga, G. M. Mller, R. W. L. Godschalk, E. J. Moonen, E. F. M. Wouters, and F. J. Van Schooten (2010, April). A profile of volatile organic compounds in breath discriminates COPD patients from controls. *Respiratory Medicine 104*(4), 557–563.

Wier, L. M., A. Elixhauser, A. Pfuntner, and D. H. Au (2011, February). Overview of Hospitalizations among Patients with COPD, 2008.