# Integrated data analysis for early warning of lung failure

Geisinger Health Collider Project: Stage 1

The Outliers: Rebecca Barter and Shamindra Shrotriya

November 7, 2015

## 1 Abstract

Chronic obstructive pulmonary disease (COPD) is a major cause of mortality worldwide, however many of those with COPD remain undiagnosed. Individuals with undiagnosed COPD typically experience related adverse health effects resulting in increased utilization of health care services. One of the primary reasons for hospitalizations amongst undiagnosed COPD patients is pneumonia; amongst patients with a secondary diagnosis of acute exacerbation of Chronic obstructive pulmonary disease (COPD), pneumonia was the primary reason for hospitalization for 22.3 percent of admissions. For this project, 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 this prediction problem focus only on utilizing clinical information, however we aim to incorporate external data sources primarily related to relevant socioeconomic factors that are not captured by the clinical records using a process called "data blending".

## 2 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. Crucially, however, it is estimated that a further 12 million adults in the U.S. have undiagnosed COPD [?]. One of the most common causes of COPD is cigarette smoking, however as many of 10-20% of COPD patients have never smoked and only a fraction of smokers develop COPD. These insights are suggestive of alternate risk factors such as genetic and environmental. With this in mind, our goal is to "blend" external environmental data with the clinical data using a novel blending approach with the hopes of increasing our ability to predict undiagnosed COPD amongst those hospitalized with pneumonia. .

## 3 Data

The clinical data we will use in this project will include data collected from patients who are a part of the Geisinger Health System. The dataset consists of data from a total population of 88,000, among whom, 1,033 have a "discharge diagnosis of COPD". The clinical dataset contains approximately 80 unique variables sourced primarily from medical records with a few variables sourced from billing records. We will be comparing two sub-populations

1. The COPD subpopulation: all patients who have been diagnosed with COPD.
2. The non-COPD pneumonia subpopulation: those who are diagnosed at a single visit with pneumonia and subsequently recover. We do not include in this subpopulation those who are re-admitted at multiple visits with pneumonia, since this may be indicative of undiagnosed COPD.

## 3.1 ”External” data sources

As a major focus of this project is to blend external data sources with the clinical data described above. This task is poised with a number of challenges. Firstly, to ensure that the imputed data resulting from the blending process (see methods below) is as transparent and believable as possible, it is important that the external data sources come from a population that is similar to our base clinical dataset from Geisinger. Notably, since all Geisinger medical centers are located in Pennsylvania, ideally, we would also like our external data to be from Pennsylvania. Suppose, for example, that we were interested in testing the effect of biomass fuel combustion exposure on COPD and for our clinical population we imputed the extent of biomass fuel combustion exposure based on data sourced from developing countries in which they use a lot more biofuels, then these imputed values will be extremely misleading. and we will describe three possible sources of data. Next, we require the data to be “publicly available”, which we interpreted as “easily accessible to Geisinger”. As a result, some of our “external” data sources include data collected by the Geisinger health system, but not included in the provided COPD clinical dataset. The primary reason for this approach is three-fold: first of all, it is extremely difficult to find individual-level data that is publicly available (presumably due to privacy concerns) AND from a trusted source (i.e. not sourced from some guy sitting in his basement making up numbers for fun); secondly, using data collected by Geisinger ensures that the underlying population mostly matches the population from which the COPD data is drawn, a key concern for the validity of the conclusions drawn from our subsequent analyses. Finally, for the publicly available individual-level data that we could find, it is extremely difficult to find adequate variables to match on. For example, in most cases, the best we could do is matching on age and gender and approximate location, however if we are aiming to impute a variable such as smoking or exposure to biomass fuel combustion, the imputed values is likely to be extremely noisy and not particularly trustworthy (for example, it is highly unlikely that all females, aged 27 who live in Danville, Pennsylvania have the same smoking status; we need more common information between the two datasets.)
The most trusted publicly available data sources we could identify are significantly less granular. For example,

### 3.1.1 Smoking data from the Geisinger psychiatric diseases dataset

As mentioned above, one of the most common causes of the

### 3.1.2 Insurance data from Geisinger health systems

Employment, medical claims (longitudinal example)

### 3.1.3 Air pollution data from Pennsylvania

### 3.1.4 Weather data from Pennsylvania

# 4 Methods

In this section, we will describe our intended methodology for both data blending as well as our indented analysis

## 4.1 Data blending

Here we will describe the data blending approach and describe methods to show that it is robust and “reliable”

## 4.2 Prediction

Our prediction approach, perhaps mention assessing causality (but only if we have a very clear question in mind). Discuss withholding a test set, and examples of the physical methods to take.

## 5 Discussion

## 6 Conclusion