Integrated data analysis for early warning of lung failure

Geisinger Health Collider Project: Stage 1

The Outliers: Rebecca Barter and Shamindra Shrotriya

November 6, 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 (occupational?) data with the clinical data using a novel blending approach with the hopes of better predicting undiagnosed COPD amongst those hospitalized with pneumonia. Our analysis will take two approaches: the first combines all data sources in an attempt to improve prediction overall, and the second identifies individuals whose COPD diagnosis is not well predicted using the clinical data and examine whether the predictive performance amongst these people can be improved using external variables.

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 and billing records. We will be comparing two sub-populations

1. The COPD subpopulation: all patients who have been diagnosed with COPD.

2. The 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.

As a major focus of this project is to blend external data sources with the clinical data described above

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