Integrated data analysis for early warning of lung failure

Geisinger Health Collider Project: Stage 1

The Outliers: Rebecca Barter and Shamindra Shrotriya

November 16, 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

3.1.1 Key Requirements - 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 quality in our external data we felt that that following strict requirements need to be satisfied for it to be effective in our case:

- Data come from a publicly availiable sources
  - Here 'publicly availible' means stored in a digital format on the internet on secure servers which are accessible globally
- Data source is well trusted in the data science/ statistical community
  - In this case 'trusted source' means that the sourced data was created/ audited by an arganisation that is well cited and relied upon in the data science/ statistical community. This would generally include data provided by (but not limited to) government departments
- Data contain no Personally Identifiable Information (PII)
  - In order to ensure that no privacy issues of any patients are breached the external data sourced must contain no information on the address, name, DOB information about a specific patient/ person (PII). For data sourced within the Geisinger internal patient database care must be taken to ensure that privacy is protected through discussions with the Geisinger team
- Data are sourced preferably from Pennsylvania to be best representative of the Geisinger population
  - 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.
  - This effectively ensures 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.
- Data contains the key fields upon which we want to "blend" on
  - The primary variables which will be used to blend data are varioues combinations age, gender, zipcode and date.
  - Ideally our dataset should have multiple combinations of these fields so that we could 'blend' them for each patient and get maximum variation between patients to be used for classification purposes

3.2 "External" data sources - main sources

Given this we searched for several external data sources that helped fit the identified requirements above. In our overall search for publicicly availible external data from Pennsylvania we found the following key issues overall:

- It is extremely difficult to find individual-level data that is publicly available (presumably due to privacy concerns of citizens) **and** from a trusted source (i.e. from a government body or an organisation that is not privately funded with a particular biased agenda)
- For the publicly available individual-level data that we could find, it is extremely difficult to find sufficient adequate variables to match on.
  - In most cases, the best data we found matched only on singular variables e.g. age, gender, zip-code (approximate location) and date.
  - However if our aim is 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 simply need a larger combination of the common blending fields to gain information from the external dataset.

The most trusted publicly available data sources we could identify are significantly less granular than required given the above issues faced. They are listed below with brief description and key issues in utilising them for the blending process:

- IRS Income Statistics Data by Zipcode
  - source: https://www.irs.gov/uac/SOI-Tax-Stats-Individual-Income-Tax-Statistics-ZIP-Code-Data-(SOI)
  - This is usefult to analyse change in income levels over time. Potentially useful as an indicator of stress (e.g. low income high unemployment trend) which could be a contributing factor to COPD.
- Pennsylvania Smoking Rankings by Zipcode
  - source: http://www.countyhealthrankings.org/app/pennsylvania/2015/measure/factors/9/data
  - This is only at zipcode level
  - Potentially useful to help distinguish patients with higher risk of COPD driven by smoking.
  - This is potentially useful if we blend using patient-level zipcode to get the maximum level of variation from this metric in our classification methodology
- Additional Income Level Data by Zipcode from Qubit Consulting
  - source: https://www.incomebyzipcode.com/search/
  - This is publicly availible but difficult to verify credibility. It may be a useful cross-validation against the IRS income statistics data as identified above
- Housing Vacancy Data in Pennsylvania
  - This is split by Gender and zipcode separately unfortunately, not as a composite gender-zipcode summary
  - There is only an overall gender distribution by zipcode provided which may be useful to help redistribute other zipcode-level metrics (e.g. population metrics) by gender.
- Climate and Temperature Data
  - This is split by Zip code and has a Temporal component for blending (by month)
  - This could be a useful to identify longitudinal trends in temperature and linking them to identifying any true pneumonia (non-COPD) cases. The hypothesis here is that colder temperatures may lead patients to suffer from pneumonia and help refine the classification

  - To source individual months may be a bit difficult and scraping would need to be done given the lack of API for downloading the data

- Census Fact Finder
  - source: http://factfinder.census.gov/faces/nav/jsf/pages/community$_facts.xhtmlHasthoroughcensusinforma$
  - There is also a distribution by Gender which could be potentially be used to re-distribute other metrics collected at the zipcode level by gender

- Flu Trends Data
  - source: https://public.tableau.com/views/s14$_1$5$_2$0150522/$Forecasts?ThisisonlyavailiblformainPAcitiesi.e$
  - Previously part of the Google Flu Trends Project (now discontinued by Google). Could be potentially used as a crude indicator of the likelihood of pneumonia alone occurring (pneumonia can occur as a complication of the flu virus).

3.2.1 Smoking data from the Geisinger psychiatric diseases dataset

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

3.2.2 Insurance data from Geisinger health systems

Employment, medical claims (longitudinal example)

3.2.3 Air pollution data from Pennsylvania

3.2.4 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".
In order to effectively blend data sources consistently, there need to be established common variables upon which the features from the datasets can be combined. We propose the following variables as the most effective method of blending external data sources to the COPD dataset provided:

1. Patient Age
2. Patient Gender
3. Patient zipcode or Hospital zipcode (less preferable)
4. Date of Patient Treatment

The above were deemed to be variables found commonly in publicly availible external data and also ensure that they are aggregated enough to ensure that we do not rely on personally identifiable information (PII) for the blending process which will help mitigate risks related to privacy breaches of patients from external data.
It should be noted that the key variables we blend on can be functions of the above variables e.g. if the external data is banded by age, we can band our original Geisinger dataset in the same age

bands as the external dataset prior to blending. This ensures greater flexibility in our blending methodology.
Having identified the external data sources of interest, we propose to integrate, or "blend", the sources with the clinical data as follows.

- Within each dataset, identify the variables that are common to the clinical data (such as age, gender, location, etc)
- Using the common variables we can join on the features from the external data using a composite-key values which are functions of the common variables e.g. joining by gender-age or zipcode-gender rather than just gender, age, zipcode separately. This depends on the granularity of the external data based on these common variables.
- In the above blending we should be careful to keep all observations in the original dataset fixed, so that we do not lose any original information provided
- In the case where the common variables in the external data are less granular than in the original Geisinger dataset we can effectively group/ band the relevant common variables in the original dataset to be consistent with the external dataset. The blending can occur on a composite key of these grouped/ banded external variables. This results in a loss of information at a patient level but may still be useful for classification purposes
- In the case where many features from the external data have missing values we can create a second version of the features which impute these missing values using the mean/ median value from other common variables which are not missing. In this sense we can increase the density of the merged dataset by relying on the overall median/ mean as a reasonable 'guess' of the required external data. Whether this approach is useful can be evaluated in terms of classification accuracy if including these features vs excluding them.

Overall the blending quality can be manually checked by taking a small number of random samples from the blended dataset and ensuring that the external data fields are merged correctly. The overall density of the blended dataset should be calculated by field to ensure that the external dataset does add sufficient non-sparse information to the original data.

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. Overall we are effectively considering a longitudinal classification problem here. As such we would rely on the following key metrics:

4.2.1 Classification Metrics

source: http://blog.dato.com/how-to-evaluate-machine-learning-models-part-2a-classification-metrics

- Overall classification accuracy
- Per-class accuracy–the average of the accuracy for each class
- Confusion Matrix
- Log-loss to guage a sense of entropy of classification and help tune model to minimise cross entropy
- Area Under Curve (AUC) and Receiver Operating Characteristic (ROC)

4.2.2 Key framework to evaluating effectieveness of the external data

5 Discussion

6 Conclusion