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

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 external data was suitable for the task at hand, we aimed to ensure that a number of requirements were met:

- **Publicly available**: We require that the data be "publicly available", which, here, means that the data is stored in a digital format on the internet on secure servers which are accessible globally.
- **Data is obtained from a trustworthy source**: The data source must be trustworthy; that is, the sourced data was created or collected by a legitimate organization, such as government departments.
- **The data must be de-identified**: The data much contain no Personally Identifiable Information (PII) in order to ensure that the privacy of all individuals whose information is stored in the dataset is de-identified (that is it does not contain information such as personal address, name and date-of-birth). 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.
- **The data is representative of the population of interest**: 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. The imputed values from this external data source will be extremely misleading and unlikely to be representative of the amount of biomass fuel combustion on similar people (e.g. similar age, gender, occupation, etc) in developed countries.
- **The data must contain some variables common to the clinical dataset**: The primary variables based on which we aim to blend data are various combinations of age, gender, zipcode and date. Ideally, to ensure that the blended dataset is as realistic as possible, our external dataset should have, as a bare minimum, each of these variables.

Although these requirements seem fairly straighforward; in practice, it turned out to be extremely difficult to find relevant publicly available individual-level data with variables common to the clinical dataset. As a result, we decided to take two alternative approaches.
First, we first propose to use data collected within 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, as noted above, it is extremely difficult to find individual-level data that is both publicly available AND from a trusted source. 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 quickly became clear that is is extremely difficult to find adequate variables to match on (where the matching is to match individuals in the external dataset to similar individuals in the clinical dataset). For example, in most cases, the best we could do is matching on age, 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).
Our second approach to identifying external data is to use the widely available location-based data (such as weather data, air pollution data or flu trends data), which cannot be imputed at an individual-level. Values of variables imputed from these types of less-granular data sources will

be the same for all patients who attend the same Geisinger medical clinic, with the presumption that these patients live within the general proximity of the medical clinic.
The following sections describe the data that we intend to blend (using various methods that will be described below), with the clinical data.

3.1.1 Smoking data from the Geisinger psychiatric diseases dataset

As mentioned above, one of the most common causes of COPD is smoking. Although the COPD dataset provided by Geisinger, does not contain the smoking status attribute, the psychiatric diseases dataset does. Assuming that it is not possible to obtain the smoking status of the COPD patients in our dataset, we could use the common attributes of the psychiatric diseases and COPD datasets to predict/impute smoking status. This is an ideal dataset for matching, since (1) the data are sourced from the same population as the patients in the COPD dataset, (2) the data are sourced from a trusted source, and (3) there will be a large number of overlap between the variables. Further, if there are common patients in each dataset, then we could use these patients to devise a supervised learning algorithm which could predict the smoking status of the non-overlapped patients in the COPD dataset (although the method we discuss in our blending section below does not assume any overlap in the two datasets).

3.1.2 Insurance data from Geisinger health systems

Write something about (1) the longitudinal flurry of medical claims/hospitalizations that preceeds COPD diagnosis - information that can be obtained from the insurance data (although this could be found in our clinical data too?), and/or (2) the employment information; biofuels inhalation idea

3.1.3 Air pollution data from Pennsylvania

Write about this too!

3.1.4 Weather data from Pennsylvania

And this!

4 Data blending

In this section, we will describe our intended methodology for both data blending for each dataset mentioned above as well as a general description of our indented analysis. Given that we have external data of various different levels of granularity, we will describe each situation individually.

4.1 Individual-level data blending

For the datasets for which we can perform individual-level matching, such as the Geisinger psychiatric diseases dataset and the insurance data, we have developed a two-stage blending method. The first stage is a matching stage which involves, for each individual in our clinical dataset, identifying individuals in the external dataset that are most similar in terms of the covariates common to both datasets. The second stage is an imputation stage which involves imputing or predicting the external variables of interest for the clinical patients by aggregating the values obtained by the matched external individuals. To set up some notation, for observation i in the clinical dataset, we can define a vector of the p observed variables by

$$\mathbf{y}_i = (y_{i1}, ..., y_{ip})^T, \qquad i = 1, ..., n$$

similarly, for the jth observation in the external dataset, we can define a vector of the q observed variables by

$$\mathbf{z}_j = (z_{j1}, ..., z_{jq})^T, \qquad j = 1, ..., m$$

Figure 1 describes this process for the first individual in the clinical dataset in a general setting. Suppose that we can reorder these variables so that the first $k \leq \min(p, q)$ variables are common to both datasets, then we can re-write our observations as

$$\text{clinical observation } i: \quad \mathbf{y}_i = (x^c_{i1}, x^c_{i2}, ..., x^c_{ik}, y_{i(k+1)}..., y_{ip})^T, \qquad i = 1, ..., n$$

$$\text{external observation } j: \quad \mathbf{z}_j = (x^e_{j1}, x^e_{j2}, ..., x^e_{jk}, z_{j(k+1)}..., z_{jq})^T, \qquad j = 1, ..., m$$

where $x_{i1}, ..., x_{ik}$ correspond to the common variables, $y_{j(k+1)}, ..., y_{jp}$ correspond to the variables that are in the clinical dataset but not in the external dataset, and $z_{i(k+1)}, ..., z_{ip}$ correspond to the variables that are in the external dataset but are not in the clinical dataset.
Our goal is to predict the unobserved values of the variables $z_{k+1}, ..., z_q$ for the patients in the clinical dataset. As mentioned above, this process is done in two steps: matching and imputation.

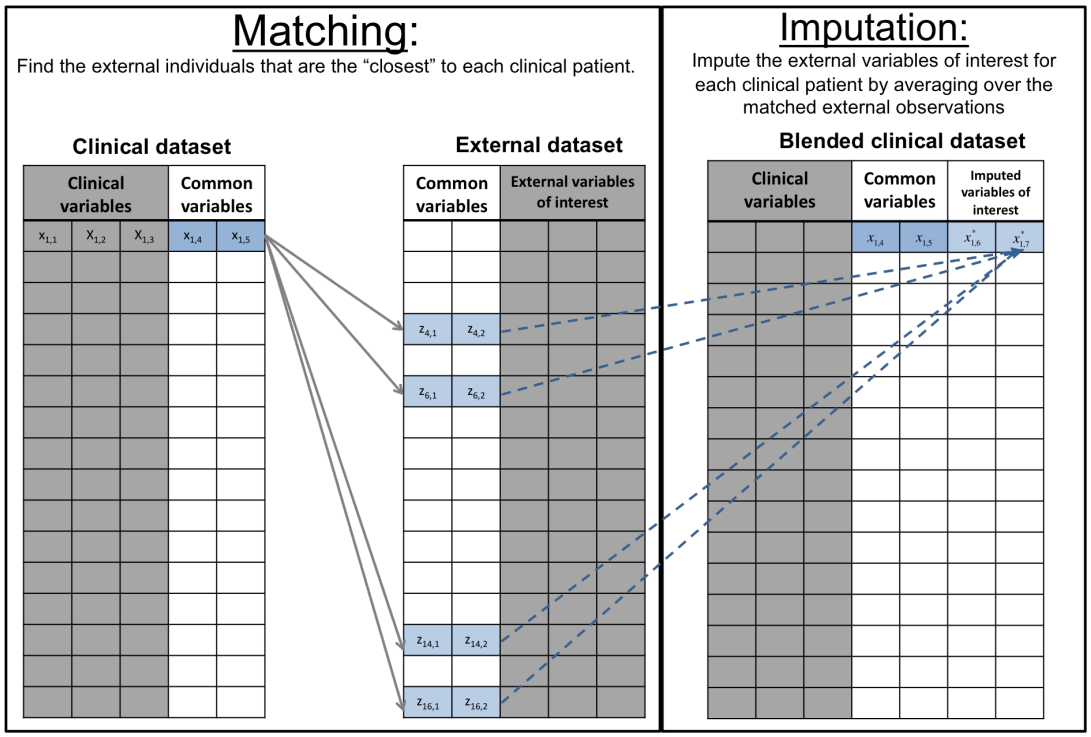

Figure 1: The data blending process for individual-level blending. For each individual in the clinical dataset, we must identify "matched" individuals in the external dataset that are the most similar based on the common variables for both datasets. Then the external variables of interest are imputed for the clinical dataset based on the matched individuals in the external dataset

4.1.1 The matching step

To perform matching, we must be able to define a measure of distance between observations in the clinical and external datasets based on the variables that are common to both. We can define the common variables for an observation in the clinical dataset by

$$\mathbf{x}^c_i = (x^c_{i1}, ..., x^c_{ik})^T$$

and the common variables for an observation in the external dataset by

$$\mathbf{x}_j^e = (x_{j1}^e, ..., x_{jk}^e)^T$$

The distance between an observation i in the clinical dataset and an observation j in the external dataset, will be defined based on the common variables only, that is

$$d(\mathbf{y}_i, \mathbf{z}_j) = d(\mathbf{x}_i^c, \mathbf{x}_j^e)$$

where we can define the distance metric, d, in a number of ways, such as the euclidean distance metric:

$$d_{euclidean}(\mathbf{x}_i^c, \mathbf{x}_j^e) = \sum_{l=1}^{k} (x_{ij}^c - x_{jl}^e)^2,$$

the Mahalanobis distance metric:

$$d_{mahalanobis}(\mathbf{x}_i^c, \mathbf{x}_j^e) = \sqrt{(\mathbf{x}_i^c - \mathbf{x}_j^e)^T S^{-1} (\mathbf{x}_i^c - \mathbf{x}_j^e)},$$

where $S = cov(\mathbf{x}_i^c, \mathbf{x}_j^e)$ is the covariance matrix, and the inverse correlation metric:

$$d_{corr}(\mathbf{x}_i^c, \mathbf{x}_j^e) = 1 - corr(\mathbf{x}_i^c, \mathbf{x}_j^e)$$

4.1.2 The imputation step

4.1.3 Assessing reliability

assess reliability: compare with if we used the obesity dataset. Comapre over subsetting and different imputation methods, etc.

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

7 Appendix

7.1 Other external data ideas

Despite these challenges this we searched for several external data sources that might fit the identified requirements above. 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`
  - 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.xhtml
  - Has thorough census information which can be searched by PA zipcode
  - 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_15_20150522/Forecasts?%3Aembed=y&%3AshowTabs=y&%3Adisplay_count=no&%3AshowVizHome=no
  - This is only availible for main PA cities i.e. Philadelphia, Pittsburgh and State College
  - 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).