Clinical Case: Predicting Asthma Healthcare Cost

Background

Asthma is a respiratory disease which causes constriction and inflammation of the airways in the lungs. The disease can often cause acute airway constrictions sometimes known as "asthma attacks." Clinicians refer to this airway restriction as "bronchospasm" and it results in cough, wheezing, shortness of breath, and extreme situations such as death from respiratory arrest. Currently, approximately 10 people die from asthma attacks per day in the US.

Asthma is most commonly diagnosed in childhood and requires a test of lung function known as spirometry to make a definitive diagnosis. Since spirometry requires patients to carefully follow specific sequences of instruction, testing may be difficult for young childen.

Often small children with asthma are instead diagnosed after they have a severe asthma attack requiring emergency care or hospital admission. Once asthma is diagnosed, it can be monitored for severity and assessed for treatment response. Treatment often consists of inhaled medications that helps combat airway constriction and inflammation.Patients that do not get timely and appropriate treatment can need avoidable emergency care and hospitalizations. These visits can add thousands of dollars per year in avoidable medical expenses to the US healthcare system. The medical costs among patients with uncontrolled asthma are 3-times higher than controlled patients.

Objective

Imagine that you are an attending physician at the UNC-health care system or a risk manager at the Blue Cross Blue Shield (BCBS). Over the last winter, the number of asthma patients suddenly increased, and the number of deaths and related costs also increased. You will want to identify the characteristics of people who are at high risk for asthma.This will help you to reduce misdiagnosis, corresponding medical expenditures, and ideally improving outcomes among these patients through applying targeted interventions. The first step is to develop a valid and reliable prediction model to identify patients most likely to require high healthcare resource utilization and high healthcare costs. Therefore, an accurate risk prediction model of future high-cost patients would allow you to more efficiently target the patients in your organization at greatest risk.

Data

You have identified the Medical Expenditure Panel Survey (MEPS)[1] as a potential data resource to help you achieve your objective. "The Medical Expenditure Panel Survey (MEPS) is a set of large-scale surveys of families and individuals, their medical providers, and employers across the United States. MEPS is the most complete source of data on the cost and use of health care and health insurance coverage."[1] Using 2009 MEPS data, you have identified a nationally representable sample of 2,375 asthma patients. The MEPS data contains more than 2,000 variables, but the dependent variable you want to predict is the

[1] https://meps.ahrq.gov/mepsweb/

total healthcare expenditure in 2009 (TOTEXP09). You can see more details at MEPS website and linked table[2]. More details will be explained during the class.

R scripts are prepared in a Jupyter Notebook environment. You can access those notebooks at: https://healthit.unc.edu:8000/. Please use the given account and password to log in. There are several scripts are prepared: (1) data processing, (2) descriptive analysis, (3) correlation analysis, and (4) regression analysis. More details about how to use those notebooks and basic introduction to R will be covered during the class.

Variables

Understanding the target phenomenon of the analysis and corresponding variables involved is the first step in data analysis. Please look at the documentation[3] and the coding book[4]. Without understanding those variable, good analysis cannot be expected. Our preliminary content analysis[2] will provide a foundation for this process.

For instance, Aspirin Exacerbated Respiratory Disease (AERD) is a medical condition consisting of three key features: asthma, chronic/recurrent rhinosinusitis (inflammations of sinuses and nasal cavity), and nasal polyps. The symptoms are a result of an abnormal reaction from the bodies immune system, known as a hypersensitivity reaction, to NSAIDs or aspirin. This is in contrast to the typical allergic response which can be caused by environmental allergens such as pollen or dust. The disorder is thought to be caused by an anomaly in the metabolism of a substance known as arachidonic acid. Medications such as NSAIDs or aspirin block the COX-1 enzyme, a critical enzyme involved in arachidonic acid metabolism. This leads to increased production of proinflammatory cysteinyl leukotrienes, a series of chemicals involved in the body's inflammatory response. This resulting overproduction cause severe exacerbations of asthma and allergy-like symptoms. Thus, taking Aspirin regularly or not might have a great influence on the total medical expenditure. Taking aspirin is just one potential variable. Many other variables should be included to build robust predictive models.

Descriptive Analysis

Now you have some confidence in the potential relationship between aspirin and asthma. However, scientific evidence is needed to support this assumption. Descriptive analysis can be used to find important variables and to provide scientific evidence. Feature is another name of variable in the domain of machine learning. Feature selection is one of the most essential step in machine learning, and descriptive analysis can support this selection. Please use the R script (descriptive analysis) to find important variables.

[2] https://docs.google.com/spreadsheets/d/1lMfRMBCCnglP-M-15OmhHdy55ARmFYDV-99Fdo1fhNY/edit?usp=sharing
[3] https://meps.ahrq.gov/data_stats/download_data/pufs/h129/h129doc.shtml
[4] https://meps.ahrq.gov/mepsweb/data_stats/download_data_files_codebook.jsp?PUFId=H129

Predictive Analysis

Finally, you are about to create a model to predict the total medical expenditure regarding asthma. There are many issues to solve. For instance, there are variables that interact with each other. In other words, some valuable patterns can be identified only by observing two or more variables at the same time. You might select too many variables which can introduce noises to the model. In this case, you might want to select predictive variables or to conduct dimensionality reduction.

You might not be familiar with the terms introduced here. Do not worry. All the concepts will be covered during the class. Please use the R script (regression analysis) to practice with predictive models.