CMSC 12300 Final Project Proposal

Hannah Diamond-Lowe, Zakir Gowani, Bonnie Fan

Datasets we'll be using:

Below are a number of datasets related to healthcare provisioning, cost of medical procedures, and Obamacare signups ranging from the period 2011 - 2013.

Treatments by areas, number of patients, cost, and average Medicare payment
https://data.cms.gov/Medicare/Inpatient-Prospective-Payment-System-IPPS-Provider/97k6-zzx3
Census data on the eligible uninsured, including demographic details:
http://marketplace.cms.gov/exploreresearch/census-data.html
Survey data demographics, sign-ups (taken from http://data.illinoishealthmatters.org/):
http://factfinder2.census.gov/faces/nav/jsf/pages/searchresults.xhtml?refresh=t
Obamacare Signups
http://acasignups.net/past-projections :
Private Health Plan Data by Quarter, from 2011 to 2013:
http://www.cms.gov/CCIIO/Resources/Data-Resources/health_plan_finder_data.html

Questions we want to explore:

1) How do regional income and demographics influence cost of care and choice of coverage plan?
2) Are Obamacare sign-up totals and rate or time of sign-up, per state, related to state attributes such as population size, political classification (red/blue/swing), GDP, employment, or previous insurance coverage?
3) How many previously uninsured Americans are now covered under the Affordable Care Act? What sort of forecasts can we make about the cost of coverage in the upcoming year following the fixed March 31 sign-up date?

How will we explore the data:

We will implement (parallelized) algorithms that act on user-customizable subsets of our data to generate summary statistics, including but not limited to: top-k (for most expensive treatments in some region, most expensive healthcare by states/zipcodes, most common reason ailments per region), multivariable linear regression (how do averages of state attributes and obamacare sign-up totals per state relate?), and random forests (given proportions of chosen healthcare plans for a state, can we predict the state's political party?). K-means clustering methods can also be used to group hospital regions or co-morbid factors to further determine treatment centers with similar characteristics of patients with similar needs, with a possible matching algorithm to combine these two sets.

Ideal outcome of first prototype:

Map visualization of cost of coverage for different conditions on a zipcode level, across the United States. Make searchable by ailment, cost, region. Include demographic information (GDP, political affiliation, rates of unemployment) to generate some summary statistics.

Hypothesis, results we expect to obtain:

We hypothesize that areas of lower GDP will correlate with higher costs of care, since regions of lower GDP are more likely to have high unemployment and therefore less health coverage. It is likely that hospitals will charge more for procedures in these areas so that patients with insurance will cover the costs of procedures for those without insurance. We also predict that lower GDP will positively correlate with higher rates of sign-up for Obamacare. With our analytic tools we will confirm or reject these hypotheses as well as explore the parameter space and see what kinds of hidden and unexpected correlations lie across our data samples.