**Keeping Medicare Costs Under Control**

**Introduction**

Medicare is a way that the United States government attempts to use social insurance to cover the costs of medical procedures for those who are 65 and older. Medicare Part B specifically covers many of the things Part A does not cover, mainly that it helps to cover health care costs of outpatient services.

The Provider Utilization and Payment Data: Physician and Other Supplier Public Use File (hereby referred to as PUPD) dataset describes how often and for how much a provider billed Medicare for various services. The dataset includes information for 880,000 participating Medicare providers and includes names, credentials and provider type. The data also contains information for 6,000 types of services as listed in the Healthcare Common Procedure System (HCPCS), and represents payment information for $77 billion paid out to Medicare providers in 2012. That amount is more than 13% of the total spending on Medicare in that year. The main dataset will also be supplemented by data from 2013, and other datasets which will help describe the problem. Every row in the PUPD dataset corresponds to a specific physician and a HCPCS code which means that every row contains the information of which procedure was performed, where it was performed, how much the physician on average billed for the procedure, and how many times the procedure was performed.

**Problem Formulation**

Medicare costs are extremely high, so even a minor percentage difference can mean millions of dollars in savings for taxpayers. Also, in the current medicare market, payers, patients, and providers have only limited access to information regarding utilization and costs. The goal of this project is to provide models to help reduce costs and optimize the system to reduce costs of all three parties involved (payer, patient, and provider).

**Objectives**

In this project, the PUPD dataset will be used to build models that helps to optimize the US healthcare system in several ways:

* Identify potential provider fraud
  + Determine outliers and estimate whether they are related to fraudulent activity
  + Identify known defrauders and use them to train machine learning models to identify possible fraudulent activity - A Machine Learning approach
  + Identify possible upcoding (Using higher paying HCPCS codes in lieu of lower paying codes) - A Clustering and Classification approach
* Locate opportunities to reduce costs
  + Optimization by locale - An Optimization approach
    - Can we determine whether it is more efficient to travel for some procedures?
    - Would providers benefit from specializing in services by locality?
  + Predicting provider charges
    - Can we use PUPD dataset to determine costs for non-Medicare patients?
* Forecasting

**Paper Summaries**

* Predicting Healthcare Fraud in Medicaid: A Multidimensional Data Model and Analysis Techniques for Fraud Detection -D. Thornton, R. Mueller, P. Schoustsen, J. Hillegersberg
  + This paper seeks to measure the different kinds of fraud, and the levels to which each person is accountable for those frauds. The multidimensional approach leads to lots of ways to catch fraud by examining the relationships that patients and providers have with each other, and also the policies behind them. Ultimately the paper is more of a meta analysis than something concrete, but some of the methods may be applicable to this project.
* Forecasting medical cost inflation rates: A model comparison approach -Q. Cao, B. Ewing, M. Thompson
  + This paper is attempting to forecast medicare expenditures using an ARMA model. While the information within is ultimately useful, as it can be utilized to forecast the ways in which different categories of spending are changing, the ARMA model will not be terrible useful, due to the fact that most of our data is not of a time series type.
* Assessment of Predictive Modeling for Identifying Fraud within the Medicare Program -S. Parente, B. Schulte, A. Jost, T. Sullivan, A. Klindworth
  + Similar to the Predicting Healthcare . . . paper, they attempt to identify fraud using a multidimensional approach. They did actually attempt to find fraudsters by looking at extreme values in various categories, and use probabilities to attempt to identify them. After identifying the fraudsters, they attempted to quantify how much money was being illegally taken from the Medicare system by each fraudster, and then aggregated those results to estimate the total lost in the United States.

**Methodology**

The project will begin with some exploratory analysis to identify statistical properties of the dataset and to try to evaluate variables of interest. It will attempt to investigate issues such as over billing and upcoding, suspicious patterns (by specialty), and to evaluate geographic anomalies.

After the initial exploration, the work will be divided up so that fraud, location optimization, and forecasting can all be performed and tested. The models will be evaluated and will be compared to other research in the areas of interest.

**Evaluation**

The evaluation of the three parts will be fairly simple. Standard measures (such as R2, AUC, etc.) will be used to evaluate the traditional models when compared to themselves, and the predictive models will utilize the 2013 data to determine their accuracy.