# Investigating Statistical Models for Medicare Claims

LOGO

Gabriel Coleman, Yian Lin, Ganesh Tilve

Faculty Advisors: Ian Duncan and Roberto Molinari

Department of Statistics and Applied Probability, University of California, Santa Barbara

LOGO

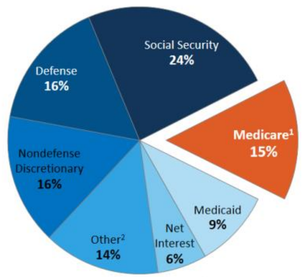
## ABSTRACT

Healthcare spending continues to rise, making cost management a priority for the healthcare industry. Our objective is to predict Medicare claims through statistical methods in order to identify patients for medical intervention. We model our data through linear regression, Generalized Linear Models (GLM), quantile regression, random forests, and Gradient Boosting Methods (GBM). Our analysis shows the traditional linear regression approach does not perform as well as the other methods, and there is significant potential to reduce healthcare expenditures by accurately predicting high cost patients.

## MEDICARE EXPENDITURE

Figure 1

**Medicare as a Share of the Federal Budget, 2015**




**Total Federal Outlays, 2015: $3.7 trillion**

**Net Federal Medicare Outlays, 2015: $540 billion**

NOTE: All amounts are for federal fiscal year 2015. ¹Consists of mandatory Medicare spending minus income from premiums and other offsetting receipts. ²Includes spending on other mandatory outlays minus income from offsetting receipts.
SOURCE: Congressional Budget Office, Updated Budget Projections: 2016 to 2026 (March 2016).

## MEDICARE COST DISTRIBUTION 2013

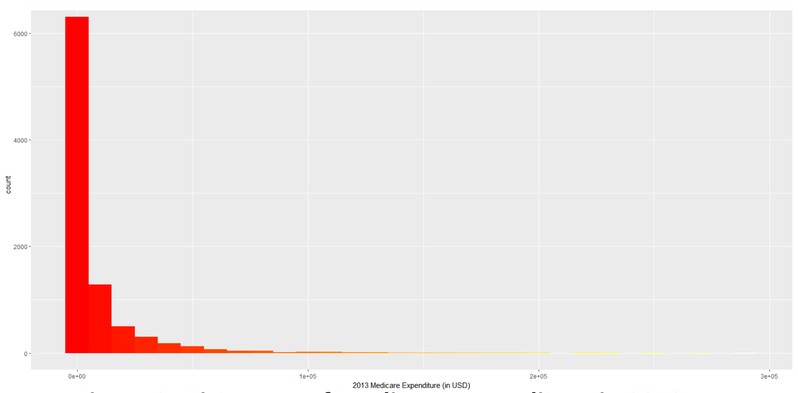



**Figure 2: Histogram of Medicare Expenditure in 2013**

## METHODOLOGY

- **Approach 1: build models to predict 2012 costs**
  - ✓Input: Patients' (demographics + health status) in Year N
  - ✓Output: Patients' costs in Year N
  - ✓Application: Provide ideas on how to control the costs after identifying the future high-cost patients
  - ✓Test set: 2013 data set
- **Approach 2: build models to predict 2013 costs**
  - ✓Input: (demographics + health status + costs) in Year N
  - ✓Output: costs in Year N
  - ✓Application: Help us identify the high-cost patients in the future year based on previous information
  - ✓Test set: 20% of observations
- **Approach 3**: Build a Logistic model to predict which patients will be high cost (greater than $15,000) in as in approach 1
  - ✓We use R to simulate interventions in 2013 patients based on predicted cost

## METHODS

- **Generalized Linear Models (GLM)**
  - Gaussian, Inverse Gaussian, Stable, Gamma
  - Binomial (Logistic)
- **Machine Learning Methods**
  - Random Forest and Gradient Boosting Method (GBM)
- **Variable Selection Methods**

| Method | Formula |
|---|---|
| **LASSO** | $\min_{\beta}\|y - X\beta\|^2 + \lambda\|\beta\|^1$ |
| **Ridge** | $\min_{\beta}\|y - X\beta\|^2 + \lambda\|\beta\|^2$ |
| **Elastic Net** | $\min_{\beta} \|y - X\beta\|^2 + \lambda[\alpha\|\beta\|^1 + (1-\alpha)\|\beta\|^2]$ |

- **Model Selection Criteria** $\frac{1}{n}\sum_{i=i}^{n}|\hat{Y}_i - Y_i|$
  - Mean Absolute Error

## RESULTS

**Approach 1:**

| Method | GLM | Random Forest | GBM |
|---|---|---|---|
| Selected Model | Inverse Gaussian (LASSO) | RF (ntrees=100, max depth=15) | GBM (ntrees=100, max depth=5) |
| MAE | 4328 | 599 | **578** |

**Table 1: Modeling results part I for approach 1**

- The selected GBM is the best model.
- Machine learning methods predict better than GLM.
- Five most important predictors of the selected Random Forest: sum of HCCs and the 4 categories of claims

**Approach 2:**

| Method | GLM | Random Forest | GBM |
|---|---|---|---|
| Selected Model | Inverse Gaussian (LASSO) | RF (ntrees=100, max depth=10) | GBM (ntrees=100, max depth=5) |
| MAE | **5788** | 5827 | 5909 |

**Table 2: Modeling results part I for approach 2**

- The selected Inverse Gaussian is the best model
- Machine learning methods and GLM have similar accuracy in predicting costs in future
- Five most important predictors of the selected Inverse Guassian: sum of HCCs and the 4 categories of claims

**Approach 3:**

- In this approach we were able to tell if patients would cost more than $15,000 at an accuracy rate of 78%
- Using our model, we achieve a savings rate of 4.5% in total expenses in 2013 for our population
- $30 Billion: The potential savings if this simulation were preformed on the entire Medicare recipient population

## CONCLUSION

- We can potentially save $30 Billion from using logistic predictive models
- Our most common variables used are inpatient claims, outpatient claims, prescription claims, age, and sum of HCC's.
- GBM and Random forest are the best predictive models