Project Proposal: Machine Learning with Medicare Data

Adelle Lin, Yuta Okazaki, and Andrew Fair

Medicare is the US’s national social health insurance program for older adults. Medicare serves approximately 50 million Americans, the majority of which are 65 and older (some younger individuals with disabilities are also served).

The Centers for Medicare and Medicaid Services (CMS) makes available a synthetic public-use dataset[[1]](#footnote-0) representing a 5% sample of 2008 Medicare beneficiaries, with their insurance claims from 2008 to 2010. This dataset is fully anonymized and “synthesized”, meaning that while the data are based upon real beneficiaries, they are altered in such a manner to make them unlinkable to actual beneficiaries (for the preservation of privacy). CMS makes this dataset publicly available, primarily for data entrepreneurs to use for application development and research training purposes. The dataset represents more than 2 million individual beneficiaries, and includes beneficiary summary information, inpatient claims, outpatient claims, carrier claims, and prescription drug events.

We propose to investigate predictive modeling with machine learning using this Medicare dataset. We will address the following research questions:

* Are there significant predictive differences when modeling outcomes by race, by sex, and by age?
* Which chronic conditions have the greatest/least impact on mortality?
* How do chronic conditions correlate with each other and with other features?

We have a variety of features and labels available to work with. Using only the beneficiary summary table, we can investigate relationships among the following:

* Predictive features:
  + Demographic factors (age, sex, race, geographic location)
  + Presence of chronic conditions (Alzheimer’s, heart failure, kidney disease, cancer, COPD, depression, diabetes, ischemic heart disease, osteoporosis, arthritis, stroke)
* Labels/outcomes:
  + Advent of additional chronic diseases
  + Death
  + Insurance payments (inpatient, outpatient, carrier)

For a somewhat more complex analysis, we can join inpatient and outpatient claims tables with beneficiary summary data in order to add more detailed factors into the model, such as specific diagnosis codes, specific procedural codes, number of inpatient stays, and length of inpatient stays. We will likely begin by exploring the beneficiary summary table, and add these more complex table joins to our analysis if time permits, as they will require significantly more data processing.

A convenience of using this dataset is that it is already divided up into 20 distinct sub-samples. This will make it easy to create training, validation, and test sets. We will use machine learning approaches such as neural network modeling, the naïve Bayes classifier, and support vector machines to explore the relationships outlined above.

Successful predictive modeling using these features and outcomes could be of use to clinical decision support systems and preventive resource allocation. For instance, if we can predict a high risk of depression in white, female patients over age 85 with a history of rheumatoid arthritis, a healthcare provider could use this information to make sure mental health support resources are available.

1. <https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/SynPUFs/DE_Syn_PUF.html> [↑](#footnote-ref-0)