Project Proposal:
An ICD-9/ICD-10 Based Hospital Readmission Prediction Model

Michael Gorczyca (mtg62)
Laura Jones (lej4)
Justice Sefas (jds464)
Ryan Vogan (rcv39)

Question: Utilizing scalable machine learning algorithms, can we further improve the predictive power in statistical models of patient re-admission?

Data Set(s): MIMIC-III data set and Nationwide Readmissions Database (NRD) data set.

Project Relevance: National health care expenditures are expected to be $3.4 trillion for 2016 with avoidable hospital readmissions expected to be $17 billion. In an attempt to lessen such costs while also improving hospital quality, several researchers have developed statistical models for objectively evaluating patient conditions. However, it is important to note that in the context of hospital readmission, most of these models exhibit poor predictive performance and are unsuitable for use in a clinical setting. To address this issue, we plan to develop a hospital readmission prediction model that has improved probabilistic calibration, discrimination, and accuracy in determining whether a patient with a given set of conditions will be re-admitted. Recent advances in scalability for model development and assessment, as well as our prior experiences working with large data sets, give us confidence that we can ultimately develop a model with better predictive performance than those currently available.

Relevance of Data Set(s): MIMIC-III comprises over 58,000 hospital admissions from 2001-2012 at Beth Israel Deaconess Medical Center (Boston, MA). It contains detailed information regarding the clinical care of its patients, from patient caregiver notes to the vital sign measurements made at a patients bedside each hour. MIMIC-III is generally recognized as a premier data set for evaluating patients who have been admitted to the Intensive Care Unit (ICU), given the large population of ICU admitted patients and the amount of data provided about such patients.

NRD comprises data from hospitals in 21 states, and has the benefit of having a large sample size - it contains data from approximately 14 million discharges each year. What makes NRD unique from other readmission based data sets is that it contains the reasons why a patient returns to a hospital for care and a hospital's costs for discharges with and without readmissions. NRD also contains a wealth of information about the patient (such as ICD-9 procedural and diagnosis codes, co-morbidities, and general demographics) that is essential for evaluating hospital systems in general.