*BMI 507: Clinical Decision Support*

*Topic: Research Design and Learning to Predict Using EHR Data*

Data: <http://mimic.physionet.org/gettingstarted/access/>

This data is provided via a license for the department.

MIMIC-III (Medical Information Mart for Intensive Care III) is a large, freely-available database comprising deidentified health-related data associated with over forty thousand patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012.

The database includes information such as demographics, vital sign measurements made at the bedside (~1 data point per hour), laboratory test results, procedures, medications, caregiver notes, imaging reports, and mortality (both in and out of hospital).

The latest version of MIMIC is MIMIC-III v1.4, which comprises over 58,000 hospital admissions for adults and neonates. Neonates have been filtered out and only one visit per patient has been kept.

**There are a total of 38,888 patients in the dataset.**

The data column/variable details can be found under Tables in MIMIC

**Goal: You are going to try to predict mortality (EXPIRE\_FLAG) with high accuracy using logistic regression, decision tree, and random forest.**

**There are 3 data tables:**

* Patients\_Admissions (MIMIC tables Patients and Admissions merged)

[1] "SUBJECT\_ID" "HADM\_ID" "INSURANCE" "MARITAL\_STATUS"

[5] "ETHNICITY" "DIAGNOSIS" "ADMISSION\_TYPE" "GENDER"

[9] "DOB" "EXPIRE\_FLAG

* Diagnoses (MIMIC table DIAGNOSES\_ICD and D\_ICD\_DIAGNOSES merged)

[1] "SUBJECT\_ID" "HADM\_ID" "ICD9\_CODE" "ROW\_ID" "SHORT\_TITLE" "LONG\_TITLE"

[7] "ICD9\_DEC"

* Procedures (MIMIC table CPT EVENTS)

[1] "SUBJECT\_ID" "HADM\_ID" "CPT\_CD" "CPT\_NUMBER"

[5] "CPT\_SUFFIX" "SECTIONHEADER" "SUBSECTIONHEADER" "Study"

\* I would filter the CPT events by their section header

**Step 1: Create Features**

Using the three data tables decide what types of diagnoses, procedures, and demographic information you want to include in your model to predict whether or not someone died during their ICU visit. Choose at least 15 features (more is better). If you would like to make a different cohort (ie. from the 38,888 choose only those with atrial fibrillation, stroke, etc.) and predict mortality from that, go ahead!

**Step 2: Divide Your Data Into Train and Test Sets**

Divide your data into 20% testing and 80% training sets. Check to see the distribution of people who died in your training and testing sets. Is it the same?

**Step 3: Train a logistic regression, decision tree, and random forest model to predict mortality.**

Produce sensitivity, specificity, PPV, and NPV. What features have high importance/significant odds ratios?