Project Letter of Intent

Ryan Quan, Frank Chen

2015-02-23

Declaration of Partnership

Frank Chen (fc2451) and Ryan Quan (rcq2102) will be collaborators on this project.

Research Question

This project will focus on building a prediction model for the onset of sepsis in the ICU, allowing one to better detect the need for prophylactic intervention within a critically ill patient population.

Generalizability of the model would only go as far as the confines of the MIMIC II Clinical Database, from the assumption that prediction power is only applicable within the same practice in which the EMR data was collected.

Description of Dataset

The data will be from the Multiparameter Intelligent Monitoring in Intensive Care Database (MIMIC II), which presents ICU patient records for approximately 25,000 adults at Boston's Beth Israel Deaconess Medical Center.

Number of Sepsis-related Cases by ICD-9 Code

```
SELECT code, count(*) AS count
    FROM mimic2v26.ICD9
    WHERE code LIKE '995.9%'
    OR code = '785.52'
    GROUP BY code
```

Number of Unique Subjects with Sepsis-related Complications

```
SELECT count(DISTINCT subject_id) AS sample_size
    FROM mimic2v26.ICD9
    WHERE code LIKE '995.9%'
    OR code = '785.52'
```

Methods

Feature Selection

Process for feature selection has not yet been decided. However, since our model is intended to predict onset of sepsis, we will restrict features to vitals collected within a specified time frame of admission to the ICU.

Patient Population

Patients included in the prediction model will consist of subjects who have acquired sepsis during their ICU stay. To avoid bias introduced by censorship, we will exclude samples who have not been in the ICU for longer than X hours, as patients will not have accrued enough data to make a risk assessment.

To avoid bias introduced by confounding medical interventions, patients with previously identified microbiology events and prescribed prophylactic treatment will also be excluded.

The Surviving Sepsis Campaign defines prophylactic treatment as "immediate intervention with pressors, antibiotics, and fluid resuscitation."

Analysis

As this is a supervised classification problem that requires some clinical interpretability, we have elected to use the following models to predict the onset of sepsis within the ICU:

- Logistic Regression
- Naive Bayes
- Decision Trees

We have taken note of pre-existing scoring systems, which we will use as benchmarks for comparison. For example, the SIRS criteria uses four simple rules to flag patients at risk for sepsis-related complications. In order for our prediction model to be useful in the clinical setting, we must at least achieve greater predictive accuracy than the SIRS criteria.

Moreover, since our goal is to detect early onset of sepsis, we ideally want our model to have high accuracy with data collected within the first 24 hours. As

such, we may elect to compare models trained on data collected at varying time intervals, e.g. 3 hours, 6 hours, and 12 hours after ICU admission.

Tools

We will be using the `glm`, `e1071`, `rpart`, and `caret` packages from CRAN for model training, testing, and validation. We may elect to validate using the `sklearn` library in Python.

Data for the analysis will be pulled from either the flat-files via a Python script or a virtual machine preloaded with a PostgreSQL database - both of which are available on PhysioNet.