The use of Electronic Health Records (EHR) over the past several years has generated a large data source that allows for development of machine learning models for early diagnosis, risk stratification, and clinical decision support. Generating gold-standard labels for the outcome (phenotyping) is critical to the process of developing a training cohort, but is often a labor-intensive process requiring manual chart review. Sepsis affects over a million patients annually and remains one of the largest contributors to mortality in the ICU, costing the healthcare system over 14 million dollars per year. In hopes of facilitating high-throughput development of predictive models, we propose an electronic phenotyping algorithm capable of retrospectively identifying sepsis cases from the EHR that attains high performance without the use of ICD-9 billing codes. Additionally we explore models that predict risk of mortality following sepsis on the basis of the derived EHR features.

### Data

The data of interest is contained within the MIMIC III database[2], an electronic health record database curated by MIT that houses de-identified demographics, vital signs, lab test results, procedures, medications, notes, imaging reports, and outcomes of 58,000 hospital admissions between 2001 and 2012 for 38,645 adults and 7,875 neonates at the Beth Israel Deaconess Medical Center. For classification, we label hospital admissions as a positive example only if sepsis occurs over the course of the admission based on the clinical criteria by Angus et. al[1]. For the purposes of survival analysis, we consider the set of admission with a positive sepsis label who also experienced a death in the hospital and define the time of death as the number of days since admission.

### Feature Engineering

We were successful at processing a large and diverse clinical database for the retrospective classification of sepsis cases, but the utility of the model is limited in that valid classification may only be made retrospectively and thus cannot be used for clinical decision support or real-time prediction. However, given that we are able to achieve relatively high performance without the use of ICD-9 codes, it may be possible to use this model to develop study cohorts with patients that may have been missed by models using only the ICD-9 codes for the outcome definition. Additionally, this same set of summary features attains modest performance at predicting the time-dependent risk of death in the hospital following sepsis, but the result is less strong than in the classification case.