# Project Proposal

Ryan Quan, Frank Chen

2015-03-16

## Predicting Sepsis in the Intensive Care Unit

### Research Question

This study will focus on developing a prediction model for the onset of sepsis in the ICU using clinical history and non-invasive physiological data obtained in the first six hours of admission. Using a large dataset of patients, routine clinical measurements obtained during initial stages of care, new imputation techniques, and data mining methodologies, the goal will be to facilitate advance warning of sepsis in the general critical care setting.

### Background

Sepsis is systemic inflammatory response syndrome (SIRS), secondary to a documented infection. Sepsis can present itself on a continuum that ranges from sepsis, severe sepsis, and septic shock, resulting in multiple organ dysfunction. The symptoms of sepsis are often non-specific and involve difficulty breathing, hypoxemia, hypoperfusion, and hypotension [1].

Although sepsis is a common condition worldwide, the current understanding of the pathophysiology of sepsis has increased substantially, and sepsis mortality has declined in the last two decades [2]. The reason for the decline may be attributed to improved supportive care and the inherent symptomatology of patients who fall prey to sepsis. On the contrary, epidemiologic data suggests that sepsis incidence is increasing [2]. New treatments and therapies have failed to demonstrate efficacy. Sepsis affects approximately 700,000 people per year, and accounts for approximately 200,000 deaths per year in the United States [3], amassing an annual cost of 16.7 billion dollars [4].

The best form of treatment is preventive treatment. Early diagnosis and appropriate therapy must be typically be delivered before laboratory test results are known, which bases the diagnosis on the co-presence of routine clinical measures. The SIRS criteria was developed in the 1991 International Sepsis Definition

Conference to address these concerns and is still commonly used in the clinical care setting to flag patients for risk of sepsis [1]. Patients who meet the SIRS criteria exhibit two or more of the following symptoms:

- Temperature > 38 degrees Celsius or < than 36 degrees Celsius
- Heart Rate > 90 bpm
- Respiratory Rate > than 20 or PaCO2 < 32 mm Hg
- White Blood Cell Count > 12,000/mm^3, < 4,000/mm^3, or > 10% bands

Unfortunately, the SIRS criteria has low discriminatory power in the intensive care unit as many critically ill patients who are not at risk for sepsis may also exhibit similar symptoms [5]. Previous studies demonstrated the poor utility of the SIRS criteria in identifying septic patients within a clinical care setting, in which SIRS exhibited both low sensitivity and specificity [6]. In the case of identifying patients at risk for sepsis, a test with poor sensitivity can be particularly harmful as false negatives may not receive the proper prophylactic care needed to prevent sepsis-related complications. As such, a high-recall prediction model (low false negatives) to identify patients with sepsis may provide benefits to caregivers in the form of an early warning system.

While previous studies have largely focused on predicting septic shock [7], few studies have focused on predicting earlier stages of the sepsis continuum. Multivariate logistic regression (Shavdia, 2007), decision trees [8], and Dynamic Bayesian Networks [9] approaches have been used to predict sepsis in the intensive care unit. However, these studies tended to use a large number of invasive measurements - such as arterial blood pressure - in their feature set, reducing generalizability. Moreover, while other studies looked at the last measurements taken before the onset of sepsis [10], few models incorporated summary statistics (mean/sd or other pairs) of clinical features in the feature set to capture the centrality and dispersion of these measurements over time. Our study attempts to synthesize and add to previous approaches by applying: a) "modern" classification methods (naive bayes, regularized logistic regression, and random forest) to potentially improve model performance, b) summary statistics to routine, non-invasive clinical features to capture information from time-series data, and c) imputation methods to avoid pitfalls due to missing data.

## Materials and Methods

### Dataset

The data will be obtained from the Multiparameter Intelligent Monitoring in Intensive Care Database (MIMIC II), a semi-public database which presents ICU patient records for approximately 25,000 adults at Boston's Beth Israel Deaconess Medical Center. As a large, diverse dataset of ICU patients, MIMIC II is appropriate for building prediction models for critically ill patient populations.

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.

### Patient Selection

This study will examine adults (age greater or equal to 18 years of age) who only have one ICU admission during hospital tenure. To avoid bias introduced by censorship, we will exclude samples who have not been in the ICU for longer than six hours, as patients will not have accrued enough data to make a risk assessment. To avoid bias introduced by confounding medical interventions [11], patients identified with microbial infections and who have undergone treatment with pressors, antibiotics, and fluid resuscitation within the first six hours will also be excluded. Of all patients included in the prediction model (~12,000), 2,119 will consist of subjects who have acquired sepsis during their ICU stay, designated by ICD-9 codes (995.91 and 995.92). To address the problem of class imbalance, we will undersample the majority class population. By removing some of the majority class so it has less effect on the machine learning algorithm. Of course, undersampling also means we risk removing some of the majority class instances which is more representative, thus discarding useful information.

## Outcomes

The outcome measure for this study will consist of subjects who have acquired sepsis during their ICU stay, as defined by International Classification of Diseases (995.91 and 995.92). While ICD-9 codes are not the most accurate "ground truth" labels due to administrative inconsistencies, we will give this particularly administration the benefit of the doubt (yes, this is terrible reasoning).

## Feature Selection

The patient feature set that we will be using for our analysis can be divided into two categories.

### Clinical History

The first set of features includes clinical history consisting of information available when the patient was first admitted into the ICU. These features include:

- demographic data (gender, age, etc.)
- medical history (SOFA score, SAPS-I score)
- basic health data (height, weight, etc.)

## Physiological Data

The second set of features include non-invasive measurements of physiological variables:

- blood pressure
- heart rate
- respiratory rate
- white blood cell count
- pulse oximetry
- shock index (heart rate / systolic blood pressure)

Other identified risk factors for sepsis also include:

- management of respiratory distress (ICD-9 code 786.09)
- hypoxemia (ICD-9 code 799.02)
- hypotension (ICD-9 codes 458 796.3)
- hypoperfusion (ICD-9 code 785.50)
- tachycardia
- elevated serum lactate (organ hypoperfusion)
- mental function is altered
- hyperventilation with respiratory alkolosis
- balance of pro-inflammatory and anti-inflammatory mediators
- effects of microorganisms (*Staphylococcus, Escherichia coli, Staphylococcus aureus, Klebsiella pneumoniae, Enterobacter sp, Acinetobacter baumannii, Pseudomonas aeruginosa, and Candida sp*)

We have yet to determine which of these (if any) will be included in our feature set since measurements must be considered "non-invasive" and "routine".

## Pre-Processing

We will standardize continuous clinical measures so they have a mean of 0 and a standard deviation of 1. Values falling outside the range will allow us to determine outliers and impossible values. Upon standardization, we will then compute summary statistics to capture the centrality and dispersion of clinical measures.

Upon preliminary inspection, we discovered that missing values comprise only a small fraction of all observations. However, simply excluding patients without the full feature matrix will lead to large reductions in the size of our dataset and potentially introduce bias into our models. As such, we will use a simple imputation approach in which mean feature values are derived from the patients' gender and age group.

### SQL Examples

#### 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'
```

## Analysis

### Model Selection

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

- Regularized/Non-Regularized Logistic Regression
- Naive Bayes
- Random Forests

We chose these classifiers as a means of comparison, in order to witness potential trade-offs in performance versus interpretability.

We will use 10-fold cross validation to select the best parameters.

### Model Assessment

Our goal is to develop a model that achieves high recall (low number of false positives). 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 recall than that of the SIRS criteria.

### Sensitivity Analysis

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 6 hours. However, since this is an arbitrary time point, we will compare the performance of 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 `glmnet`, `e1071`, `svm` `randomForest`, and `caret` packages from CRAN for model training, testing, and validation.

## References


1 Levy MM, Fink MP, Marshall JC*et al.* 2001 SCCM/ESICM/ACCP/ATS/SIS international sepsis definitions conference. *Intensive Care Med* 2003;**29**:530–8. doi:10.1007/s00134-003-1662-x

2 Stevenson EK, Rubenstein AR, Radin GT*et al.* Two decades of mortality trends among patients with severe sepsis: A comparative meta-analysis. *Crit Care Med* 2014;**42**:625–31. doi:10.1097/CCM.0000000000000026

3 Hartog CS, Brunkhorst FM, Bloos F*et al.* Practice of volume therapy in patients with severe sepsis: Results from a nationwide sepsis prevalence study. *Intensive Care Med* 2009;**36**:553–4. doi:10.1007/s00134-009-1736-5

4 Carrigan SD, Scott G, Tabrizian M. Toward resolving the challenges of sepsis diagnosis. *Clin Chem* 2004;**50**:1301–14. doi:10.1373/clinchem.2004.032144

5 Martin GS. Sepsis, severe sepsis and septic shock: Changes in incidence, pathogens and outcomes. *Expert Rev Anti Infect Ther* 2012;**10**:701–6. doi:10.1586/eri.12.50

6 Jaimes F, Garcés J, Cuervo J*et al.* The systemic inflammatory response syndrome (SIRS) to identify infected patients in the emergency room. *Intensive Care Med* 2003;**29**:1368–71. doi:10.1007/s00134-003-1874-0

7 Ho JC, Lee CH, Ghosh J. Septic shock prediction for patients with missing data. *ACM Trans Manage Inf Syst* 2014;**5**:1:1–1:15. doi:10.1145/2591676

8 Thiel SW, Rosini JM, Shannon W*et al.* Early prediction of septic shock in hospitalized patients. *J Hosp Med* 2010;**5**:19–25. doi:10.1002/jhm.530

9 Gultepe E, Green JP, Nguyen H*et al.* From vital signs to clinical outcomes for patients with sepsis: A machine learning basis for a clinical decision support system. *J Am Med Inform Assoc* 2014;**21**:315–25. doi:10.1136/amiajnl-2013-001815

10 Tang CHH, Middleton PM, Savkin AV*et al.* Non-invasive classification of severe sepsis and systemic inflammatory response syndrome using a nonlinear support vector machine: A preliminary study. *Physiol Meas* 2010;**31**:775. doi:10.1088/0967-3334/31/6/004

11 Paxton C, Niculescu-Mizil A, Saria S. Developing predictive models using electronic medical records: Challenges and pitfalls. *AMIA Annu Symp Proc* 2013;**2013**:1109–15./href\protect\T1\textbraceleftthttp://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900132/\protect\T1\textbraceright\protect\T1\textbraceleftthttp://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900132/\protect\T1\textbraceright (accessed 19 Feb2015).