# CSE8803: Project Draft - Septic shock prediction

Jeffrey Skonhovd

2016-12-4

## 1 Abstract

Sepsis is an extremely serious condition that occurs in emergency rooms across the country. In the United States, we know that seven hundred and fifty thousand patients develop severe sepsis and septic shock each year. This condition can quickly become fatal unless an aggressive treatment is undertaken. In this paper, we developed a machine learning application to help predict the onset of septic shock.

## 2 Introduction & Background

In the class, we have completed various rigorous programming assignments using big data technologies like Apache Hadoop and Apache Spark. My goal is to attempt to use my knowledge obtained in this course and apply it to solving a real world health care application. My application for this project will be a application that can predict Septic shock in an emergency room environment.

In this paper, we will build on solid foundation of classroom experience using Apache Spark and SciKit Learn. We will conducted a rigorous project execution and design study in which we examine all aspects of a data science desgin In addtion, we conducted an literature search to expand our mind and current knowledge.

### 2.1 Problem

Sepsis is a condition that arises when the body's response to infection injuries its own tissues and organs. (1) Sepsis is an extremely serious condition that occurs in emergency rooms across the country. In the United States, we know that seven hundred and fifty thousand patients develop severe sepsis and septic shock each year. (2) This condition can quickly become fatal unless an aggressive treatment is undertaken.

### 2.2 Motivation

Our motivation is help healthcare professional's determine which patients are at risked to develop this condition. This application will be a early warning system designed to accurately determine if a patient will develop sepsis. We will using machine learning and big data applications to complete this goal, which I will examine in depth in another section.

## 3 Project Execution

The specfic goal in this paper is to complete our project execution. The project execution is the main planing phase of the project. We have various tasks that we will need to complete. We will first gather our data for the project. This includes obtaining any permissions required. We also will design our study, which includes developing the patient cohort and the target features for this project. We will need to clean and process the data, which will be done completely in Apache Spark.

This project execution truely is the development of a simple modeling pipeline, which we use to iterate and develop our results. We will need to develop Machine Learning models for our features. In addition, we will need to use performance metrics to draw conclusions about our results.

It will be of serious importance to complete the first iteration of our project execution very quickly. We do not want to lose momentum for this project. But also, we want to quickly develop a solid foundation for this project. My goal in this paper is to focus on a simple model, and once that model is working attempting to expand on it.

### 3.1 Gather Data

In our project, we will use the MIMIC III database. 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. (6) This freely available data makes the project possible, and we are fortunate to have it. We will give a brief overview before further examination of the project execution.

MIMIC III consists of 26 tables of information provided in csv files. We can identify patients by the SUBJECT_ID in the PATIENTS table, which provides static information on each patient. We can also identify patient's hospital stay by looking at there HADM_ID on the ADMISSIONS table. If we wanted to look at individual ICU stays for a patient, we can examine the ICUSTAYS_ID on ICUSTAYS table.

Our project will focus on relating patients to events that occur at the ICU. Our basic anaylsis relies on this relationship. In order to create these relationships, we must be able to map the patients from the PATIENTS table to ICU events. We will look at the following tables for ICU events.

| TABLES |
|---|
| DIAGNOSES_ICD |
| CHARTEVENTS |

These tables are from both the hospital system, and the ICU collection systems. I developed all my features and prediction window from these two tables, which does have a couple issues. In retrospect, I hoped to get the input and output events tables into my project. I simply didn't have the time to get them implemented. Without, the input events and the output events you don't see the fluid resuscitation or fluid output, which are important features and most certainly effected my results.

## 4 Study Design

We now have a better understanding of the MIMIC III database and have defined a couple important tables, we can start designing the study and implementing parts of our project. The design study really focuses on some major milestones in the project, which include the cohort construction, defining our prediction target, and developing the correct set of features to model our problem. Addtionaly, we will examine the construction of our validation test sets.

### 4.1 Cohort Construction

Cohort construction is an important step in the design study process. We will here identify the study population for the project from the existing information contained in the MIMIC III database. Since we already have the information from historical sources like MIMIC III, we can defined this as a retrospective cohort study.

When we identify patients, we want to only examine patients who are exposed to the risk of developing septic shock. In Henry (2), they defined the following criteria for admission to the population study.

| Cohort Construction Criteria |
|---|
| Age 15 or older based on dob at ICU staytime |
| Must have at least one occurrence of GCS |
| Must have at least one occurrence of BUN |
| Must have at least one occurrence of Hematocrit |
| Must have at least one occurrence of Heart rate |

We had the task to implement these criteria into our application. The following are the set of ITEMID's I used to implement the cohort construction. As you can see, this is an rather large list but each itemid is critical to getting our cohort construction correct. We determined if a patient was part of our cohort if they had at least one entry in the ChartEvents table for these ITEMIDs.

| ITEMID | LABEL | LINKSTO |
|---|---|---|
| 211 | Heart Rate | chartevents |
| 3494 | Lowest Heart Rate | chartevents |
| 220045 | Heart Rate | chartevents |
| 220046 | Heart rate Alarm - High | chartevents |
| 220047 | Heart Rate Alarm - Low | chartevents |
| 813 | Hematocrit | chartevents |
| 3761 | Hematocrit (35-51) | chartevents |
| 227017 | Hematocrit_ApacheIV | chartevents |
| 220545 | Hematocrit (serum) | chartevents |
| 226540 | Hematocrit (whole blood - calc) | chartevents |
| 226762 | HematocritApacheIIValue | chartevents |
| 226761 | HematocritApacheIIScore | chartevents |
| 1162 | BUN | chartevents |
| 781 | BUN (6-20) | chartevents |
| 5876 | bun | chartevents |
| 3737 | BUN (6-20) | chartevents |
| 8220 | Effluent BUN | chartevents |
| 227000 | BUN_ApacheIV | chartevents |
| 227001 | BunScore_ApacheIV | chartevents |
| 225624 | BUN | chartevents |
| 198 | GCS Total | chartevents |
| 227011 | GCSEye_ApacheIV | chartevents |
| 227012 | GCSMotor_ApacheIV | chartevents |
| 227013 | GcsScore_ApacheIV | chartevents |
| 227014 | GCSVerbal_ApacheIV | chartevents |
| 220739 | GCS - Eye Opening | chartevents |
| 228112 | GCSVerbalApacheIIValue (intubated) | chartevents |
| 223900 | GCS - Verbal Response | chartevents |
| 223901 | GCS - Motor Response | chartevents |
| 226755 | GcsApacheIIScore | chartevents |
| 226756 | GCSEyeApacheIIValue | chartevents |
| 226757 | GCSMotorApacheIIValue | chartevents |
| 226758 | GCSVerbalApacheIIValue | chartevents |

In addition to Henry (2), we examined additional papers on the septic shock prediction problem including Desautels (3), and Tsoukalas (5) hoping to identify additional criteria for identifying the study population. However, we keep the criteria from Henry for our implementation.

## 4.2 Prediction Target

We already have an high level idea of what our prediction target should be. We want to identify if an patient will develop septic shock or not. But we need to a time table for our prediction! We will accomplish nothing if we do not predict a postive entry in time to perform the necessary medical procedure.

Henry (2) used a survival model called a Cox proportional hazards model using the time until the onset of septic shock as the supervisory signal. I do not have an background in statistical survival models necessary to implement one in Scala. Therefore, we must construct a specfic timetable in order to develop a prediction. For our prediction target, we will set an fix time of 24 hours before we make the prediction in order to treat this problem as a typical binary classification problem.

Now that we have the time period before we attempt to make an diagnosis, we will examine what constitutes an postive prediction. Our postive entries are determined by the following criteria.

| Positive Prediction Criteria |
|---|
| ICD-9 codes indicating infection |
| 2 SIRS Indicators |
| Sepsis Related Organ Disfunction |

### 4.2.1 Systemic inflammatory response syndrome (SIRS)

These are very common criteria in an emergency room environment, but still a little tricky to detect in MIMIC. We developed the collection of ITEMIDs for the four major sets of SIRS criteria. These sets

are Temperature, Heart Rate, Respiratory Rate, and White Blood Cell count. The associated tables are below for each set.

1. Temperature

| ITEMID | LABEL | LINKSTO |
|---|---|---|
| 676 | Temperature C | chartevents |
| 677 | Temperature C (calc) | chartevents |
| 678 | Temperature F | chartevents |
| 679 | Temperature F (calc) | chartevents |
| 223761 | Temperature Fahrenheit | chartevents |
| 223762 | Temperature Celsius | chartevents |

2. Heart Rate

| ITEMID | LABEL | LINKSTO |
|---|---|---|
| 211 | Heart Rate | chartevents |
| 220045 | Heart Rate | chartevents |

3. Respiratory Rate

| ITEMID | LABEL | LINKSTO |
|---|---|---|
| 618 | Respiratory Rate | chartevents |
| 619 | Respiratory Rate Set | chartevents |
| 224688 | Respiratory Rate (Set) | chartevents |
| 224689 | Respiratory Rate (spontaneous) | chartevents |
| 224690 | Respiratory Rate (Total) | chartevents |
| 220210 | Respiratory Rate | chartevents |

4. White Blood Cell Count

| ITEMID | LABEL | LINKSTO |
|---|---|---|
| 1127 | WBC (4-11,000) | chartevents |
| 861 | WBC (4-11,000) | chartevents |
| 1542 | WBC | chartevents |
| 220546 | WBC | chartevents |

#### 4.2.2 ICD 9 - Infection

The second of our prediction criteria, we the indication of infection in the patient. We collected this information from Angus (7). The ICD9 code information exists on DIAGNOSES_ICD. The following are our criteria of ICD9 codes for infection, I only provided an sample.

| ICD | DES |
|---|---|
| 001 | Cholera |
| 002 | Typhoid/paratyphoid fever |
| 003 | Other salmonella infection |
| 004 | Shigellosis |
| 005 | Other food poisoning |
| 008 | Intestinal infection not otherwise classified |
| 009 | Ill-defined intestinal infection |
| 010 | Primary tuberculosis infection |
| 011 | Pulmonary tuberculosis |
| 012 | Other respiratory tuberculosis |
| 013 | Central nervous system tuberculosis |
| 014 | Intestinal tuberculosis |
| 015 | Tuberculosis of bone and joint |
| 016 | Genitourinary tuberculosis |
| 017 | Tuberculosis not otherwise classified |
| 018 | Miliary tuberculosis |
| 020 | Plague |
| 021 | Tularemia |
| 022 | Anthrax |
| 023 | Brucellosis |
| 024 | Glanders |
| 025 | Melioidosis |
| 026 | Rat-bite fever |
| 027 | Other bacterial zoonoses |
| 030 | Leprosy |
| 031 | Other mycobacterial disease |
| 032 | Diphtheria |
| 033 | Whooping cough |

#### 4.2.3 Sepsis Related Organ Disfunction

1. Systolic Blood Pressure $<$ 90 mmHg;

| ITEMID | LABEL | LINKSTO |
|---|---|---|
| 6 | ABP [Systolic] | chartevents |
| 51 | Arterial BP [Systolic] | chartevents |
| 442 | Manual BP [Systolic] | chartevents |
| 3313 | BP Cuff [Systolic] | chartevents |
| 224167 | Manual Blood Pressure Systolic Left | chartevents |
| 227243 | Manual Blood Pressure Systolic Right | chartevents |
| 220050 | Arterial Blood Pressure systolic | chartevents |

2. Lactate $>$ 2.0 mmol/L;

| ITEMID | LABEL | LINKSTO |
|---|---|---|
| 818 | Lactic Acid(0.5-2.0) | chartevents |
| 1531 | Lactic Acid | chartevents |
| 225668 | Lactic Acid | chartevents |

3. creatinine $>$ 2.0 mg/dL without the presence of chronic dialysis or renal insufficiency

| ITEMID | LABEL | LINKSTO |
|---|---|---|
| 791 | Creatinine (0-1.3) | chartevents |
| 5811 | urine creatinine | chartevents |
| 3750 | Creatinine (0-0.7) | chartevents |
| 1525 | Creatinine | chartevents |
| 220615 | Creatinine | chartevents |

Chronic Dialysis or renal insufficiency are indicated by an ICD-9 code of V45.11 or 585.9.

4. bilirubin $>$ 2 mg/dL without the presence of chronic liver disease and cirrhosis

| ITEMID | LABEL | LINKSTO |
|---|---|---|
| 4948 | Bilirubin | chartevents |
| 225651 | Direct Bilirubin | chartevents |
| 225690 | Total Bilirubin | chartevents |

Chronic liver disease and Cirrhosisas are indicated by an ICD-9 code of 571 and any of the subcodes.

5. international normalized ratio (INR) $>$1.5;

| ITEMID | LABEL | LINKSTO |
|---|---|---|
| 815 | INR (2-4 ref. range) | chartevents |
| 1530 | INR | chartevents |
| 227467 | INR | chartevents |
| 220561 | ZINR | chartevents |

## 4.3 Feature Construction

In this section, we will go over our process for feature construction. in this process, we will determine our diagnosis date, our prediction window, our index data, and the observation window for our pipeline.

The diagnosis date is the date the target outcome will occur. These is the first date that one of our prediction criteria is true. We will traverse all of the possible events, and determine the earliest date. The events for ICU data are in the following tables.

| Diagnosis Event Tables |
|---|
| CHARTEVENTS |
| DIAGNOSES_ICD |

We will be able to get the exact date and time information from the CHARTEVENTS table. The diagnosis date is one of the more important values in feature construction.

We will use the diagnosis date to determine our prediction window and index date. Since in a previous section, we determined that we will treat our model like an tradtional classification problem. We did this by setting up an fixed 24 hour window before we attemped our predication. The 24 hour period before our diagnosis date is our prediction window. Our index date will be defined as 24 hours before our diagnosis date.

The observation window is an important part of feature construction. The observation window for this project will be defined as the time between the date of the first ICU entry and the index date. We can select any event as a feature if those event entries fall between that date.

We used our Sepsis Related Organ Disfunction criteria to select the index date. The first entry in the ChartEvents table, should be a strong criteria for the onset of Sepsis. However, I discovered in practice that this can occur very early, which eliminates a lot of entries.

## 4.4 Feature Selection

Feature Selection is probably the most important step in our design study and the most subjective. We are lucky in the fact that there are a lot of different events that can be determined as features in this project. Our goal in feature selection is to discover the most truly predictive features in the model.

For this paper, we wanted to keep the feature model simple at first and use the example features from Henry (2). In this paper, they develop a TREWScore in order to help predict septic shock. Well, in the additional materials provide in that paper they provide us sample feature coefficients. These sample feature coefficients are a great starting point for continuing feature selection.

| Sample Features | Feature Description |
|---|---|
| GCS | Glasgow coma score |
| BUN | BUN |
| creatinine | |
| Areterial PH | Blood pH as measured by an arterial line |
| Platelets | Platelet count in the bloodstream |
| SBP | Systolic blood pressure |
| RR | Respiratory rate |
| PaO2 | Partial pressure of arterial oxygen |
| HR | Heart rate |
| FiO2 | FiO2 |
| WBC | White blood cell count |

These features are all implemented as averages over the entire stay in the ICU during the prediction window. The following are the ITEMIDs I used to implement these features.

| ITEMID | LABEL | LINKSTO |
|---|---|---|
| 198 | GCS Total | chartevents |
| 1162 | BUN | chartevents |
| 781 | BUN (6-20) | chartevents |
| 5876 | bun | chartevents |
| 3737 | BUN (6-20) | chartevents |
| 225624 | BUN | chartevents |
| 791 | Creatinine (0-1.3) | chartevents |
| 3750 | Creatinine (0-0.7) | chartevents |
| 1525 | Creatinine | chartevents |
| 220615 | Creatinine | chartevents |
| 780 | Arterial pH | chartevents |
| 828 | Platelets | chartevents |
| 6 | ABP [Systolic] | chartevents |
| 51 | Arterial BP [Systolic] | chartevents |
| 442 | Manual BP [Systolic] | chartevents |
| 455 | NBP [Systolic] | chartevents |
| 492 | PAP [Systolic] | chartevents |
| 666 | Systolic Unloading | chartevents |
| 3313 | BP Cuff [Systolic] | chartevents |
| 3319 | BP PAL [Systolic] | chartevents |
| 3325 | BP UAC [Systolic] | chartevents |
| 7643 | RVSYSTOLIC | chartevents |
| 6701 | Arterial BP #2 [Systolic] | chartevents |
| 224167 | Manual Blood Pressure Systolic Left | chartevents |
| 227243 | Manual Blood Pressure Systolic Right | chartevents |
| 225309 | ART BP Systolic | chartevents |
| 618 | Respiratory Rate | chartevents |
| 619 | Respiratory Rate Set | chartevents |
| 224688 | Respiratory Rate (Set) | chartevents |
| 224689 | Respiratory Rate (spontaneous) | chartevents |
| 224690 | Respiratory Rate (Total) | chartevents |
| 220210 | Respiratory Rate | chartevents |
| 490 | PAO2 | chartevents |
| 779 | Arterial PaO2 | chartevents |
| 211 | Heart Rate | chartevents |
| 3494 | Lowest Heart Rate | chartevents |
| 220045 | Heart Rate | chartevents |
| 220046 | Heart rate Alarm - High | chartevents |
| 220047 | Heart Rate Alarm - Low | chartevents |
| 1040 | BIpap FIO2 | chartevents |
| 1206 | HFO FIO2: | chartevents |
| 189 | FiO2 (Analyzed) | chartevents |
| 190 | FiO2 Set | chartevents |
| 191 | FiO2/O2 Delivered | chartevents |
| 3420 | FIO2 | chartevents |
| 3422 | FIO2 [Meas] | chartevents |
| 1863 | HFO-FiO2 | chartevents |
| 2518 | HFO- FIO2 | chartevents |
| 2981 | FiO2 | chartevents |
| 1127 | WBC (4-11,000) | chartevents |
| 861 | WBC (4-11,000) | chartevents |
| 4200 | WBC 4.0-11.0 | chartevents |
| 1542 | WBC | chartevents |
| 220546 | WBC | chartevents |
| 4948 | Bilirubin | chartevents |
| 225651 | Direct Bilirubin | chartevents |
| 225690 | Total Bilirubin | chartevents |

### 4.5 Validation

Now, we have finished the heavy lifting in the project. We need to develop a sensible validation method. In this paper, we use a simple 70% training and 30% test to split our test and training cases. We always preformed an shuffle before the split.

### 4.6 Data Processing

Now that we have completed the study design, we can focus on the implementation of the project. We most over come various obstacles in order to implement this program in Apache Spark. We proccess the data on an EC2 Amazon Web Services cluster using Apache Spark for the ETL portion of the experiment. We have peformed similar operations in our third homework. However, I determined that it would be best if we saved our features as libsvm files. This was to let us use machine learning algorithms outside the feature construction phase of our Spark program.

### 4.7 Modeling Pipeline

The modeling pipeline is an continuation of the design study in the previous section. The steps of an appropriate modeling pipeline iteration are in the following table.

| Modeling Pipeline |
|---|
| Predictive Target |
| Cohort Construction |
| Feature Construction |
| Feature Selection |
| Predictive Model |
| Performance Evaluation |

We have already spent considerable energy on the first four steps in the modeling pipeline, so we will now focus on our predictive models and our plan for performance evaluation.

We used machine learning models that are already built into SciKit Learn Libray. This will allow us to have quick integration. This will allow us to focus on our models, and not the implementation of them. We determine that logistic regression and decision trees would be the best models to start with.

We wish to anaylsis the performance of our machine learning models. The model metrics will be standard machine learning metrics which we have covered in this course. We will use AUC, Accuracy, Precision, Recall and F-Score to evaluate our models. These are implemented in SciKit.

## 5 Results

### 5.1 Performance Metrics & Kaggle

The following section will show our final performance metrics for our selected features. This section will show our various performance metrics on our three different machine learning models. We appear to perform well on our algorithms for our metrics, but we found different issues when we moved to Kaggle.

#### 5.1.1 Performance Metrics

| Model | Accuracy | AUC | Precision | Recall | F-Score |
|---|---|---|---|---|---|
| Logistic Regression | 0.91037402964 | 0.520986114114 | 0.701754385965 | 0.0438596491228 | 0.0825593395253 |
| SVM | 0.908357697349 | 0.504601136251 | 0.6 | 0.00986842105263 | 0.0194174757282 |
| Decision Tree | 0.912591995161 | 0.550785914075 | 0.649006622517 | 0.107456140351 | 0.184383819379 |

#### 5.1.2 Kaggle

We did not perform well on Kaggle. My final submission was 49% correct, as of now the bottom. I believe I have a bad feature set, but I must also have an bad cohort selection. I only had features for 650 of the provided patients. I defaulted to '0', when I couldn't make an prediction.

# 6 Conclusion

## 6.1 How well do our results match the Paper?

We did not perform well in this project. My index date was certainly wrong. My features are certainly to similar between the control and the cases of Septic Shock.

## 6.2 Any pitfalls observed trying to replicate the experiment from the paper?

I had trouble getting started. I certainly spent to much time on the design phase of the project. I wish I identified the most critical features at first.

Spark is a bit tricky to configure, but on the sample data I had no issue. But the slow run times did hold me back, and I would certainly consider them a pitfall.

# 7 References


- (1) `https://en.wikipedia.org/wiki/Sepsis`
- (2) K. E. Henry, D. N. Hager, P. J. Pronovost, and S. Saria. A targeted real-time early warning score (TREWScore) for septic shock. Science Translational Medicine, 7(299):299ra122–299ra122, Aug. 2015.
- (3) Desautels T, Calvert J, Hoffman J, Jay M, Kerem Y, Shieh L, Shimabukuro D, Chettipally U, Feldman MD, Barton C, Wales DJ, Das R Prediction of Sepsis in the Intensive Care Unit With Minimal Electronic Health Record Data: A Machine Learning Approach JMIR Med Inform 2016;4(3):e28 URL:`http://medinform.jmir.org/2016/3/e28` DOI: 10.2196/medinform.5909 PMID: 27694098
- (4) Paxton C, Niculescu-Mizil A, Saria S. Developing Predictive Models Using Electronic Medical Records: Challenges and Pitfalls. AMIA Annual Symposium Proceedings. 2013;2013:1109-1115.
- (5) Tsoukalas A, Albertson T, Tagkopoulos I. From Data to Optimal Decision Making: A Data-Driven, Probabilistic Machine Learning Approach to Decision Support for Patients With Sepsis JMIR Med Inform 2015;3(1):e11 URL: `http://medinform.jmir.org/2015/1/e11` DOI: 10.2196/medinform.3445 PMID: 25710907 PMCID: 4376114
- (6) MIMIC-III, a freely accessible critical care database. Johnson AEW, Pollard TJ, Shen L, Lehman L, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, and Mark RG. Scientific Data (2016). DOI: 10.1038/sdata.2016.35. Available at: `http://www.nature.com/articles/sdata201635`
- (7) Angus, Linde-Zwirble, Lidicker, Clermont, Carcillo, Pinsky. Epidemiology of severe sepsis in the United States: Analysis of incidence, outcome, and associated costs of care. Crit Care Med (2001).