Prediction of Sepsis and In-Hospital Mortality Using Electronic Health Records

Article *in* Methods of Information in Medicine · September 2018

DOI: 10.3414/ME18-01-0014

CITATIONS
9

READS
1,521

5 authors, including:

Anahita Khojandi
University of Tennessee
52 PUBLICATIONS **152** CITATIONS
SEE PROFILE

Xueping Li
University of Tennessee
127 PUBLICATIONS **1,384** CITATIONS
SEE PROFILE

Rebecca Koszalinski
University of Oklahoma Health Sciences Center
20 PUBLICATIONS **59** CITATIONS
SEE PROFILE

Prediction of Sepsis and In-Hospital Mortality Using Electronic Health Records

A. Khojandi[1*], V. Tansakul[1], X. Li[1], R.S. Koszalinski[2], W. Paiva[3]

1 Department of Industrial and Systems Engineering, University of Tennessee, 851 Neyland Drive, John D. Tickle Building, Knoxville, TN 37996-2315, United States

2 College of Nursing, University of Tennessee, 1200 Volunteer Blvd., Knoxville, TN 37996, United States

3 Center For Health Systems Innovation, Oklahoma State University, Business Building Stillwater, Oklahoma 74078-4011, United States

Summary:

Objectives. Our goal was to develop predictive models for sepsis and in-hospital mortality using electronic health records (EHRs). We showcased the efficiency of these algorithms in patients diagnosed with pneumonia, a group that is highly susceptible to sepsis. **Methods.** We retrospectively analyzed the Health Facts® (HF) dataset to develop models to predict mortality and sepsis using the data from the first few hours after admission. In addition, we developed models to predict sepsis using the data collected in the last few hours leading to sepsis onset. We used the random forest classifier to develop the models. **Results.** The data collected in the EHR system is generally sporadic, making feature extraction and selection difficult, affecting the accuracies of the models. Despite this fact, the developed models can predict sepsis and in-hospital mortality with accuracies of up to 65.26±0.33% and 68.64±0.48%, and sensitivities of up to 67.24±0.36% and 74.00±1.22%, respectively, using only the data from the first 12 hours after admission. The accuracies generally remain consistent for similar models developed using the data from the first 24 and 48 hours after admission. La[1]stly, the developed models can accurately predict

[*] corresponding author: khojandi@utk.edu

sepsis patients (with up to 98.63±0.17% accuracy and 99.74%±0.13% sensitivity) using the data collected within the last 12 hours before sepsis onset. The results suggest that if such algorithms continuously monitor patients, they can identify sepsis patients in a manner comparable to current screening tools, such as the rule-based Systemic Inflammatory Response Syndrome (SIRS) criteria, while often allowing for early detection of sepsis shortly after admission. **Conclusions.** The developed models showed promise in early prediction of sepsis, providing an opportunity for directing early intervention efforts to prevent/treat sepsis.

Keywords: Predictive analytics, sepsis, in-hospital mortality, electronic health records

1. Introduction

Sepsis is the systemic inflammatory response to severe infection, typically pneumonia, gastrointestinal or urinary tract infection [2], and can cause serious consequences for patients. The mortality rate following sepsis can reach up to 30%, with 50% and 80% for severe sepsis and septic shock, respectively [2]. Once a patient develops sepsis, the mortality rate goes up when left untreated. Therefore, detection of high-risk patients is necessary in order to decrease mortality through early intervention and optimal care.

Because sepsis is a system inflammatory response to infection, it is generally associated with elevated heart rate, temperature, and respiratory rate, as well as either low or high white blood cell (WBC) count. Accordingly, healthcare providers currently rely on patients' physiological symptoms to identify sepsis cases [3]. For instance, Systemic Inflammatory Response Syndrome (SIRS) criteria, which was introduced in 1992, categorizes a patient as septic from having two or more of the symptoms presented in Figure 1 [3]. In 2016, Sepsis-3 was introduced to replace the SIRS criteria with a new risk-stratification tool. In Sepsis-3, sepsis is defined as life-threatening organ dysfunction caused by a dysregulated host response to infection [4]. Quick Sequential Organ

Failure Assessment (qSOFA) was also introduced within Sepsis-3 to be used with patients who have suspected infection and are likely to have prolonged stay in intensive care units (ICUs) or to expire in the hospital [4]. The validation of Sepsis-3 and also qSOFA are subjects of ongoing research [5]. Identifying septic patients using these recent definitions and assessment tools is somewhat complex, which coupled with the lack of requisite data, may not be practical in our dataset [5]. Hence, in this study we opted to use the well-established SIRS criteria.

2. Objectives

The goal of this study was to retrospectively analyze historical electronic health records (EHR) data to develop models that can predict sepsis and in-hospital mortality. Specifically, we used a robust classification technique, namely random forest [15], on physiological data collected shortly after admission to predict future incidence of sepsis and in-hospital mortality, and to draw insights about the changes in patient symptoms. In addition, we used random forest on the physiological data collected shortly leading to incidence of sepsis to examine the efficiency of the algorithm in distinguishing sepsis patients. In general, these models can help healthcare practitioners in early detection of sepsis and provide patients with timely, personalized treatments before a sharp increase in the risk of developing sepsis or in-hospital mortality.

The time of sepsis is generally not recorded in EHR systems. Hence, in this study, we categorized patients as septic as soon as they meet the well-accepted SIRS criteria. In addition, in this study we limited our attention to adult patients diagnosed with pneumonia, a group that is highly susceptible to sepsis.

3. Literature Review

There exists an extensive body of work on the use of data-driven models to predict sepsis or mortality. Taylor et al. [6] used emergency department (ED) visits data of patients age of 18 or

older and developed sepsis as meeting SIRS criteria with infectious admitting diagnosis to predict in-hospital mortality by using random forest model, classification and regression tree (CART) model, logistic regression model, and previously developed clinical decision rules (CDRs). Their results show that random forest outperformed other models and had the highest area under the receiver operating characteristic (ROC) curve. Gultepe et al. [7] used EHR of adult patients who met a minimum of two on SIRS criteria and were admitted through the emergency department to build models using support vector machine (SVM) and Bayesian network (BN) to predict lactate level and mortality. These models were trained for sepsis patients, and all patients regardless of sepsis status, and achieved accuracies of up to 72.8% and 71.5% in predicting mortality, respectively. Giuliano et al. [8] used Project IMPACT dataset of adults with an admitting ICU diagnosis of sepsis to assess the predictive value of early detection of sepsis using physiological data recommended by the Surviving Sepsis Campaign (SSC). They obtained an accuracy of approximately 62% in predicting sepsis using the logistic regression algorithm. Another study [9] used data from the ICU to detect sepsis in real-time using decision trees, support vector machines, and Naïve Bayes (NB) algorithms. All developed models successfully detected all patients experiencing severe sepsis and septic shock, except for the NB algorithm that misclassified only one septic shock patient as a severe sepsis patient, resulting in an accuracy of 99.82%.

4. Methods

4.1 Dataset

We used the data pulled from the Health Facts® (HF) dataset [1]. The de-identified dataset was provided by the Center for Health Systems Innovations (CHSI) at Oklahoma State University. The dataset contains EHRs from approximately 490 hospitals under Cerner Corporation, collected over approximately 14 years. The dataset includes the details of patients' demographics (e.g., gender,

age, marital status, race), patients' information (e.g., admitted information, discharged information), clinical events (e.g., vital signs), lab procedure results (e.g., WBC count), medications administered (e.g., name of medication, order strength of medication), and diagnosis information (e.g., ICD-9 code [10]). Specifically, here we only focused on data from 2008 to 2015 on adult patients (18 years or older) who were admitted due to either physician or clinical referral, and were diagnosed with pneumonia, captured by the ICD-9 code [10]. We used Structured Query Language (SQL) to select the appropriate data from this dataset for analysis. Table 1 summarizes the demographics of patients who were diagnosed with pneumonia and admitted with referrals.

4.2 Data Cleansing

As in most clinical datasets, our EHR dataset contained a large number of null values and duplicated observations, especially under patient encounters, clinical events, and lab procedure results tables. For this analysis, null observations were removed and among duplicated observations with the same date and time, the one with the larger value was kept.

Because the data were collected from multiple institutions over a long time-span, there were major inconsistences across the units in which the data were reported, especially for vital signs and WBC count. Hence, we converted the data when necessary. In addition, not all data was clinically meaningful after unit conversion. Hence, we removed the entries that fell outside of the following ranges: A respiratory rate between 4 and 60 breaths per minute, temperature between 32.2 and 41.1 degrees Celsius, heart rate between 30 and 200 beats per minute [11], and WBC count between 500 and 50,000 cells/μL [12].

4.3 Response Variables and Features

We used two response variables in this study, namely, in-hospital mortality and sepsis. For in-hospital mortality, we used the discharge description which was recorded under the patient

encounters table. We eliminated observations corresponding to "not mapped" and "unknown," as well as null values. As mentioned in Section 2, the EHR did not contain the time of sepsis if it was developed. Hence, we used the well-established SIRS criteria [3] to estimate the time of sepsis if it occurred. Specifically, we retrospectively examined each patient encounter to determine whether they acquired sepsis and if so, collected its initiation time.

We used a total of 57 features, including both categorical and continuous variables as shown in Table 2. Categorical variables include demographics information, such as gender, race, payer code, and age groups [13]. Continuous variables included information on vital signs, namely, heart rate, respiratory rate, and temperature, as well as WBC count. We calculated the basic statistics such as minimum, maximum, and standard deviation, as well as information entropy, particularly Shannon entropy [14], for all of the four continuous variables.

In addition, when possible, we generated features based on changes in consecutive clinical events. That is, we calculated the *differences in consecutive values*, as well as the *proportional differences*. Specifically, the differences in conservative values were calculated by finding the difference between consecutive observations for each vital sign and WBC count. The proportional differences were calculated when dividing the differences in conservative values from vital signs and WBC count by the differences in time between these consecutive observations. We then calculated the basic statistics of these features.

Note that generating the differences in consecutive values and the proportional differences features requires at least two values. Hence, for some patients, due to the low frequency of data collection for some features, such as WBC count, differences in consecutive values and the proportional differences features could not be calculated. Therefore, we performed the analysis in two ways: (1) Kept the differences in consecutive values and the proportional differences features

and removed patients from the dataset with fewer than two entries for vital signs or WBC count, and (2) removed the differences in consecutive values and the proportional differences features and kept the patients for whom the parameter could not be calculated in the dataset.

4.4 Experiments

We performed three main experiments as follows:

Experiment I: Used the EHR data from the first 12, 24, and 48 hours after admission to predict which patients would develop sepsis;

Experiment II: Used the EHR data from the first 12, 24, and 48 hours after admission to predict which patients would expire;

Experiment III: Used the EHR data from the 12, 24, and 48 hour-windows leading to sepsis to predict which patients would develop sepsis.

We used two feature subsets as follows:

(a) All features

(b) All features except differences in consecutive values and proportional differences

For each experiment, the dataset was refined to only include patients who had length of stay (LOS) longer than the number of hours used in the corresponding analysis. For instance, in Experiment I, when predicting which patients would acquire sepsis using the EHR data from the first 12 hours after admission, we excluded patients who acquired sepsis within the first 12 hours. Note that in Experiments I and II, we used the data from a time-window immediately after admission to determine whether or not patients would develop the condition of interest, i.e., sepsis or mortality, respectively. In Experiment III, however, we used a different approach to prepare the dataset. For the sepsis patient subgroup, we used the data from a time-window leading to sepsis onset, including the data points collected at the time of sepsis onset (as marked by SIRS criteria). For the non-sepsis

patient subgroup, we used the data from a randomly selected time-window of the same size during patients' stay. We collected the data from the two subgroups in a dataset, along with their corresponding response variables of sepsis/non-sepsis, and used the dataset in Experiment III.

4.5 Classification Algorithm and Performance Metrics

In this study, we used random forest classifier [15] to develop the models. Random forest is an ensemble learning method and can be used in classification and regression problems. Random forest relies on the aggregate results from a series of decision trees. We particularly used random forest in this study as the algorithm is very robust against overfitting due to randomly selecting subset of features at each split as it grows decision trees [15]. We used Scikit-Learn library [19] in Python 2.7 [18] for implementation. We partitioned the dataset into 85% and 15% for training and test sets. Based on the results of our preliminary experiments, we opted out of tuning hyper-parameters for random forest models using a validation set to not reduce the sizes of the training and test sets.

Our training datasets were, in general, highly unbalanced with respect to the response variables, e.g., there were approximately nine times more instances of expired patients than non-expired patients in the cleansed dataset under Experiment II with feature subset (a). To ensure that the developed models did not favor the more represented observations in the dataset, we used the downsampling technique to generate a series of balanced sub training datasets from the initial training dataset in which both classes were represented with the same proportion. In addition, we exploited warm-starting to achieve higher accuracy. Warm-starting in random forests is a technique implemented to iteratively add trees in a forest as opposed to fitting a whole forest at once. Doing so, the algorithm reuses the solutions of previous iterations, which generally helps improve the training process of the model. Specifically, for random forest, we developed a 700-

tree forest by building one tree at a time on a new sub training set, while applying warm-starting technique, and aggregated them into one model. Finally, the best trained models were applied on the corresponding, separate test sets to objectively evaluate the performance of the models.

For all experiments, we report sensitivity, specificity, accuracy, F1 score, the area under the ROC curve (AUC), and Matthews correlation coefficient (MCC) for the test sets. Accuracy gives the proportion of predicted values that match the true response values and F1 score is a weighted average of precision and recall. AUC gives the probability that for any randomly selected pair of positive and negative observations (e.g., sepsis and non-sepsis patients, respectively), the model will rank the positive observation higher than the negative one. Lastly, MCC provides a measure of the overall performance of the model. MCC ranges between -1 and 1, where 0 represents no better than random prediction.

4.6 Ethical Considerations

The study was performed in compliance with the World Medical Association Declaration of Helsinki on Ethical Principles for Medical Research Involving Human Subjects, and was reviewed by the Institutional Review Board (UTK IRB–17– 04148–XM).

5. Results

In our dataset, the average LOS of patients who developed sepsis was 179.14 hours, compared to 53.96 hours for the average LOS of patients who did not develop sepsis. Figure 2 presents the breakdown of the dataset with respect to meeting SIRS criteria along with the discharge description. Consistent with our experiments, in the figure we stratify patients based on their LOS, i.e., LOS more than 12, 24 and 48 hours, as well as meeting SIRS criteria. We report the raw numbers and percentage of patients in each subcategory. For instance, out of the total of 332,006 patients remained in the dataset after cleansing, 261,258 (or approximately 79%) have a LOS that

is greater than 12 hours, out of which 106,938 (41%) acquired sepsis at some point. Approximately 27% of patients with a LOS greater than 12 hours, acquired sepsis after 12 hours, i.e., 73% of patients acquired sepsis within the first 12 hours after admission. This highlights the importance of predicting/detecting sepsis immediately, or within only a few hours, after admission.

Next we present the results of the three experiments and examine the impact of the choice of feature set and balancing on the results. To obtain more stable average results and also investigate the variability of the results, all analyses are executed 10 times and the mean performance and 95% confidence intervals (CIs) across these 10 runs are reported. The results are reported in the following format: $\bar{x} \pm (t \text{ critical value}) \frac{s}{\sqrt{10}}$ with 9 degrees of freedom (df).

Table 3 summarizes the results of Experiment I, i.e., predicting sepsis, using feature sets (a) and (b). As seen in Table 3, for feature set (a), the accuracy and F1 score using the data collected within the first 12 hours after admission are approximately 63.27±1.04% and 64.26±0.81%, respectively. The accuracy remains at around 59-64% for larger time windows; however, F1 score decreases by approximately 14%. Recall that given a time window to use for predicting future onset of sepsis, we ensure that patients who develop sepsis within this time window are excluded from the experiment. Hence, the decrease in F1-score for 24-hour and 48-hour time windows may be due to the smaller datasets that are used for training the models. As seen in Table 3, the model performance slightly improves with the feature set (b), compared to feature set (a). This again may be attributed to the loss of observations when using feature set (a).

Table 4 presents the average confusion matrices across 10 runs, and summarize the results, for the models built using balanced and unbalanced training sets for Experiment I using the data from the first 12 hours after admission with feature subset (b). This table highlights the impact of

balancing the datasets in improving the model performance, e.g., F1 score increases from 62% to 64% and sensitivity increases from 60% to 67% due to balancing the training set.

Table 5 summarize the results of Experiment II, i.e., predicting mortality, using feature sets (a) and (b). As seen in Table 5, for feature set (a), the best accuracy is on the order of 70%. Consistent with the observations from Experiment I, models built using feature set (b) slightly outperform those developed using feature set (a) with respect to accuracy. In these models, F1 score is consistently low, mainly due to the highly unbalanced dataset with respect to the response variable 'in-hospital mortality.' Recall that there were approximately nine times more instances of expired patients than non-expired patients in the cleansed dataset under Experiment II with feature subset (a). Table 6 presents the average confusion matrices across 10 runs, and summarize the results, for the models built using balanced and unbalanced training sets for Experiment II using the data from the first 12 hours after admission with feature subset (b). As seen in the table, the F1 score and sensitivity dramatically increase due to balancing; however, specificity and accuracy decrease.

Lastly, Table 7 presents the results of Experiment III, i.e., predicting sepsis using the data collected in the time windows leading to sepsis. Compared to previous experiments, in this experiment the accuracy and F1 score are extremely high, at approximately 99%. This implies that the algorithm can perform in a comparable manner to the rule-based SIRS criteria to predict sepsis onset. As seen in the table, the accuracy and F1 score decrease in window size as a longer window size introduces more uncertainty to the model.

Lastly, to further investigate the generalizability and validation of our approach, we stratified the patient records based on hospital sizes and examined the accuracy of sepsis prediction using the data from the first 12 hours after admission. Specifically, we used three categories of

hospital sizes (i.e., small hospital: < 200 beds, medium hospital: 200-499 beds, large size hospitals: 500+ beds) and developed and tested models within and across these categories. When testing within the same category, we used an 80%/20% split for training and testing. Table 8 presents the mean performance and 95% CIs across 10 runs. As seen in the table, the model performances are generally consistent across hospital sizes, suggesting that models are overall generalizable.

6. Discussions, Limitations and Future Work

In the future, it is foreseeable that clinicians will be able to rely on algorithms to predict sepsis/mortality using the data collected immediately after admission. Such algorithms would then enable clinicians to intervene in a timely manner to reduce patients' risks of acquiring sepsis or an untimely death. Our results suggest that in general algorithms can predict sepsis using widely collected EHR data in a comparable manner to the rule-based SIRS criteria (Experiment III), while often being capable of sepsis prediction shortly after admission (Experiment I). Additionally, in cases where patients are facing a life-limiting illness or injury, algorithms can predict mortality using the widely collected EHR data shortly after admission (Experiment II). Predicting mortality can empower patients and their caregivers with patient-centric pain management, emotional and spiritual support, and hospice care when appropriate.

In Experiments I and II, we developed models using two feature subsets and compared the model performances. Our results showed that the models generally became more accurate when more data were available for training. For instance, although generating features, such as differences in consecutive values and proportional differences in vital signs and WBC counts, make clinical sense when it comes to detecting sepsis/mortality, including them in the model reduced the number of observations and hence, overall reduced the model performance. Adoption of automated high frequency data collection systems at bedside, which can store more data points

for patients, would allow for better examination of features such as differences in consecutive values and proportional differences in terms of their contribution to model performance.

The accuracy of the developed sepsis prediction models using the data from the first 12 hours after admission is approximately 65%; however, this accuracy gets close to 99% when accounting for the last 12 hours leading to sepsis. Hence, it is plausible to assume that the algorithms can help identify at-risk patients as early as 12 hours after admission and continue to increase in their accuracy if patients start to deteriorate or their risk of sepsis goes up in time. Therefore, these algorithms can provide much value if they are used in an online fashion, and can possibly detect at-risk patients well before rule-based techniques such as SIRS.

The developed models also allow for identifying the most important contributing factors to sepsis/mortality prediction. In this study, feature importance is determined by the total decrease in the Gini impurity criterion [21]. Table 9 presents the top ten most important features (across the 10 runs) for some of the best performing random forest models of Experiments I-III. As seen in Table 9, the entropy of respiratory rate consistently had the highest importance in discriminating sepsis/non-sepsis patients and expired/non-expired patients in both Experiments I and II when using the data from the first 12 hours after admission. We also obtained similar results when differences in consecutive values and proportional differences features were present when using feature set (a). Different from Experiments I and II, in Experiment III, i.e., predicting sepsis using the data collected in the time windows leading to sepsis, the maximum of heart rate was identified as the most important contributing factor. Further exploring the top ten most important features across the 10 runs reveal that these features are more or less consistent across the models, suggesting that the models are generally very robust and consistent, and the features identified as most important are generally reliable.

Sepsis is an important clinical event, the onset of which should be recorded in EHR systems. Under sepsis-1 and sepsis-2 definitions, patients who have infections and meet two or more symptoms under SIRS criteria [3] could be identified as septic. However, the true onset of sepsis for patients may only be identified by clinicians at bedside. Similar to most EHR systems, the system that had contributed to our dataset did not contain the diagnosis time of sepsis. Therefore, we used SIRS criteria to retrospectively approximate the time of sepsis in a given group of patients who had already been identified to have infection (i.e., patients with pneumonia). We believe recording the time of sepsis diagnosis by healthcare providers would prove very helpful in building more accurate predictive models in the future.

Note that sepsis definitions were published multiple times within 25 years, which indicates that the knowledge of sepsis is still limited. Sepsis-3 definition introduced new criteria, qSOFA and SOFA. We opted to use SIRS criteria in this study as Sepsis-3 would require keeping track of six parameters to determine whether patient encounters develop sepsis or not. Using the current dataset to mark sepsis patients with SOFA criteria would have resulted in a much smaller dataset with far fewer valid patient encounters.

We lost many patient encounters due to erroneous data or missing labels (mainly, sepsis or mortality). In our exploratory analysis, we encountered major inconsistences in units, many clinically non-meaningful values, missing labels, as well as duplicated observations in patients' information and clinical events. It is likely that most of these erroneous data or missing labels were caused by data entry error. A more careful approach to form design and/or adopting automated data collection systems would reduce such errors and help with future algorithm developments.

We acknowledge that the demographics used in this study were not diverse. The summary of demographics is shown in Table 1. The predominate race of the majority of patient encounters

was Caucasian. Further studies need to be performed to examine whether the risk factors or models translate well for other populations.

7. Conclusions

In this study, we developed models to predict sepsis and in-hospital mortality using EHR data. The developed models showed promise in early prediction of sepsis, possibly providing an opportunity for directing early intervention efforts to prevent/treat sepsis. However, the accuracy of models are expected to improve with better data collection and possibly with adoption of automated data acquisition systems that can collect and store high frequency data without direct clinician intervention. Our results suggested that having more observations in general help increase the model performance. However, further analyses are needed to determine the exact trade-off between the number of observations and the added value of more sophisticated feature engineering. Lastly, based on our results, it is clear that the algorithms can help identify at-risk patients as early as 12 hours after admission. This accuracy increases dramatically as patients are recognized to be sepsis (as marked by SIRS criteria). Hence, it is plausible that continuous monitoring of patients using these algorithms can add value by early prediction of sepsis patients and pave the way for a streamlined and improved care process.

8. Acknowledgements

This research project was conducted using the data pulled from the Cerner Corporation's Health Facts database of electronic medical records provided by XXX. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Cerner Corporation.

Conflicts of Interest

The authors declare that they have no conflicts of interest in the research.

References

1. DeShazo JP, Hoffman MA. A comparison of a multistate inpatient EHR database to the HCUP Nationwide Inpatient Sample. BMC health services research. 2015;15(1):384.

2. Jawad I, Lukšić I, Rafnsson SB. Assessing available information on the burden of sepsis: global estimates of incidence, prevalence and mortality. Journal of global health. 2012;2(1).

3. Bone RC, Balk RA, Cerra FB, Dellinger RP, Fein AM, Knaus WA, Schein RM, Sibbald WJ. Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. Chest. 1992;101(6):1644-55.

4. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, Bellomo R, Bernard GR, Chiche J-D, Coopersmith CM. The third international consensus definitions for sepsis and septic shock (sepsis-3). Jama. 2016;315(8):801-10.

5. Marik PE, Taeb AM. SIRS, qSOFA and new sepsis definition. Journal of thoracic disease. 2017;9(4):943.

6. Taylor RA, Pare JR, Venkatesh AK, Mowafi H, Melnick ER, Fleischman W, Hall MK. Prediction of In-hospital Mortality in Emergency Department Patients With Sepsis: A Local Big Data–Driven, Machine Learning Approach. Academic Emergency Medicine. 2016;23(3):269-78.

7. Gultepe E, Green JP, Nguyen H, Adams J, Albertson T, Tagkopoulos I. From vital signs to clinical outcomes for patients with sepsis: a machine learning basis for a clinical decision support system. Journal of the American Medical Informatics Association. 2014:315-25.

8. Giuliano KK. Physiological monitoring for critically ill patients: testing a predictive model for the early detection of sepsis. American Journal of Critical Care. 2007;16(2):122-30.

9. Gonçalves JM, Portela F, Santos MF, Silva Á, Machado J, Abelha A. Predict sepsis level in intensive medicine–data mining approach. Advances in Information Systems and Technologies: Springer; 2013. p. 201-11.

10. International Classification of Diseases, Ninth Revision (ICD-9) [Internet]. Centers for Disease Control and Prevention [updated September 1, 2009; cited 2017 October 14]. Available from: http://www.cdc.gov/nchs/icd/icd9.htm.

11. Bleyer AJ, Vidya S, Russell GB, Jones CM, Sujata L, Daeihagh P, Hire D. Longitudinal analysis of one million vital signs in patients in an academic medical center. Resuscitation. 2011;82(11):1387-92.

12. Barron HV, Harr SD, Radford MJ, Wang Y, Krumholz HM. The association between white blood cell count and acute myocardial infarction mortality in patients≥ 65 years of age: findings from the cooperative cardiovascular project. Journal of the American College of Cardiology. 2001;38(6):1654-61.

13. Hu G, Baker SP. An explanation for the recent increase in the fall death rate among older Americans: a subgroup analysis. Public health reports. 2012;127(3):275-81.

14. Shannon CE. A mathematical theory of communication. ACM SIGMOBILE Mobile Computing and Communications Review. 2001;5(1):3-55.

15. Breiman L. Random forests. Machine learning. 2001;45(1):5-32.

16. Haykin S, Network N. A comprehensive foundation. Neural Networks. 2004;2(2004):41.

17. Tu JV. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. Journal of clinical epidemiology. 1996;49(11):1225-

18. Oliphant TE. Python for scientific computing. Computing in Science & Engineering. 2007;9(3).

19. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research. 2011;12(Oct):2825-30.

20. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M, editors. TensorFlow: A System for Large-Scale Machine Learning. OSDI; 2016.

21. Tan PN, Steinbach M, Kumar V. Introduction to data mining. India: Pearson, 2006.

SIRS Criteria

Meet two or more of the followings:

- Temperature: $< 36℃$ or $> 38℃$
- Heart rate: > 90 beats per minute
- Respiratory rate: > 20 breaths per minute
- White blood cell count: $< 4{,}000$ cells per mm^3, $> 12{,}000$ cells per mm^3, or $> 10\%$ immature (band) forms

qSOFA Criteria

Meet two or more of the followings:

Respiratory rate: ≥ 22 breaths per minute
Altered mental status
Systolic blood pressure: ≤ 100 mm Hg

Figure 1. SIRS and qSOFA criteria.

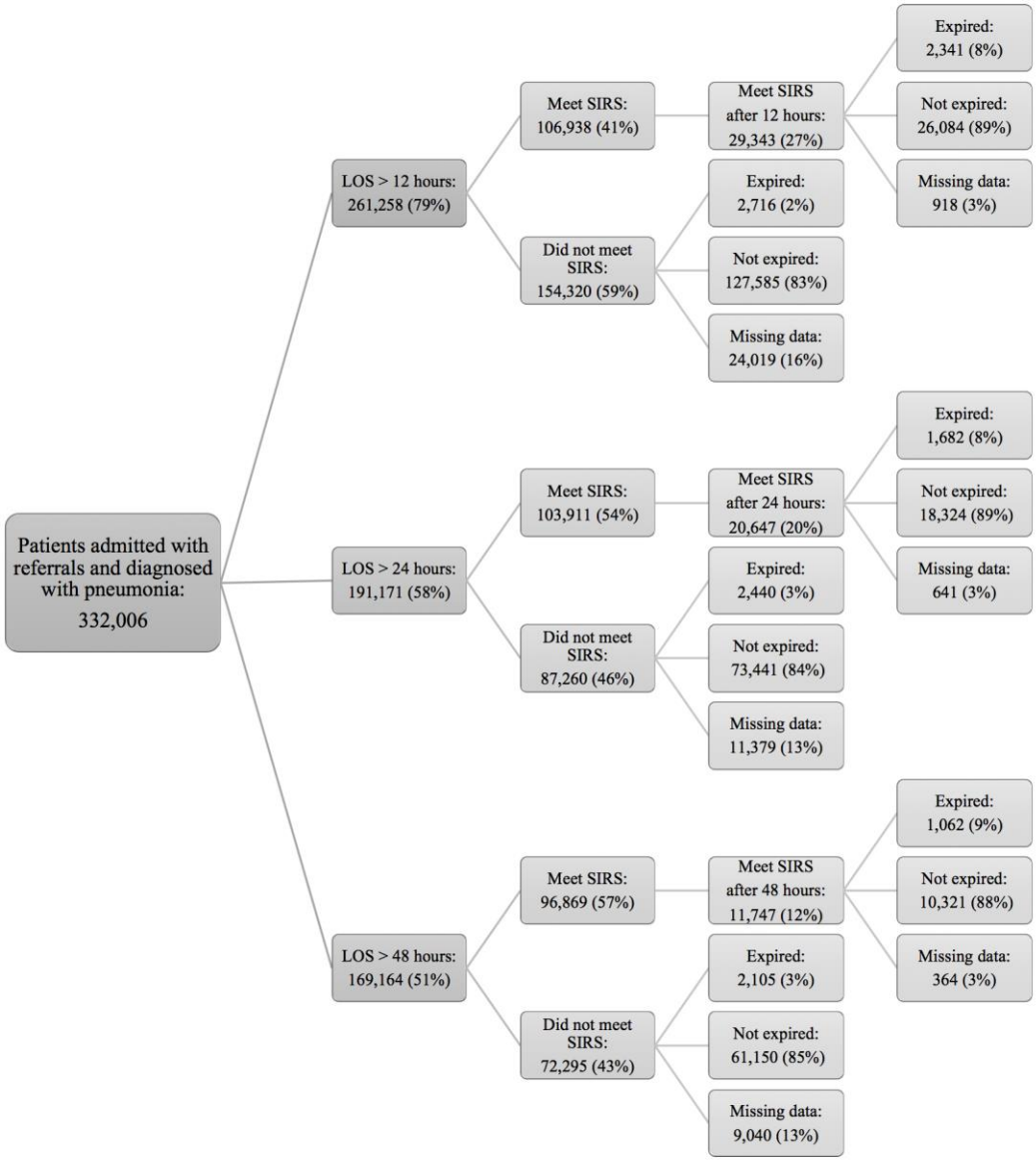

Figure 2. The breakdown of patient encounters with respect to meeting SIRS criteria and the discharge description.

Table 1: Summary of demographics.

Total patient encounters (*n*)	332,006
Gender	
Female	52.73 %
Male	47.25 %
Race	
Asian or Pacific Islander	1.22 %
African American	14.66 %
White	79.36 %
Other/ Unknown	4.76 %
Marital Status	
Married	43.83 %
Widowed	18.84 %
Single	22.58 %
Divorced	11.88 %
Unknown	2.87 %
Payer Code	
Private/HMO	20.49 %
Medicaid	7.88 %
Medicare	46.32 %
Self-pay/uninsured	5.10 %
Other	20.21 %
Age (years)	63.56±18.41

Length of Stay (hours)	97.43±739.62 (Q_1:14,Q_2:51,Q_3:127)

Table 2: Features used in predictive models.

Demographics	Age groups: 18-44 years old, 45-64 years old, ≥65 years old
	Gender: Male, Female
	Race: Asian or Pacific Islander, African American, White, Other
	Marital status: Married, Widowed, Single, Divorced, Unknown
	Payer code: Private/HMO, Medicaid, Medicare, Self-pay/uninsured, Other
Features below were applied to heart rate, respiratory rate, temperature, and WBC count	
Basic statistics	Minimum, maximum, mean, standard deviation
Signal information	Shannon Entropy
Differences in consecutive values	Minimum, maximum, mean, standard deviation
Proportional differences	Minimum, maximum, mean, standard deviation

Table 3: Mean performance and 95% CIs across 10 runs for Experiment I with feature sets (a) and (b).

Feature Set	**Data Collection Window**	**Accuracy**	**F1 score**	**Sensitivity**	**Specificity**	**AUC**	**MCC**
(a)	First 12 hours after admission	63.27% ±1.04%	64.26% ±0.81%	62.86% ±1.17%	63.82% ±2.12%	0.63 ±0.01	0.27 ±0.02
	First 24 hours after admission	64.06% ±0.69%	60. 96% ±0.67%	63.78% ±0.72%	64.38% ±0.98%	0.64 ±0.01	0.28 ±0.01
	First 48 hours after admission	59.05% ±0.59%	50.16% ±0.74%	63. 45% ±1.50%	56.96% ±1.01%	0.60 ±0.01	0.19 ±0.01
(b)	First 12 hours after admission	65.26% ±0.33%	64.40% ±0.28%	67.24% ±0.36%	63.47% ±0.56%	0.65 ±0.00	0.31 ±0.01
	First 24 hours after admission	62.21% ±0.45%	57.30% ±0.36%	61.56% ±0.42%	62.59% ±0.74%	0.62 ±0.00	0.24 ±0.01
	First 48 hours after admission	60.31% ±0.75%	49.50% ±0.68%	62.32% ±0.97%	59.33% ±0.98%	0.61 ±0.01	0.20 ±0.01

Table 4: Average confusion matrices across 10 runs for Experiment I with feature set (b) and using the data from the first 12 hours after admission. Average number of false negatives/positives and true negative/positives are rounded to whole numbers.

		Predicted+	**Predicted-**	
Balanced training set	**Condition +**	515	251	**Accuracy** ≈ 65% **F1 score** ≈ 64% **Sensitivity** ≈ 67% **Specificity** ≈ 63% **AUC** ≈ 0.65 **MCC** ≈ 0.31
	Condition -	319	553	
Unbalanced training set	**Condition +**	456	310	**Accuracy** ≈ 66% **F1 score** ≈ 62% **Sensitivity** ≈ 60% **Specificity** ≈ 71% **AUC** ≈ 0.65 **MCC** ≈ 0.31
	Condition -	252	620	

Table 5: Mean performance and 95% CIs across 10 runs for Experiment II with feature sets (a) and (b).

Feature Set	**Data Collection Window**	**Accuracy**	**F1 score**	**Sensitivity**	**Specificity**	**AUC**	**MCC**
(a)	First 12 hours after admission	64.55% ±0.40%	30.37% ±0.52%	74.00% ±1.22%	63.47% ±0.40%	0.69 ±0.01	0.24 ±0.01
	First 24 hours after admission	67.80% ±0.42%	28.63% ±0.30%	74.89% ±1.11%	67.18% ±0.56%	0.71 ±0.00	0.25 ±0.00
	First 48 hours after admission	69.03% ±0.31%	25.70% ±0.18%	69.74% ±0.45%	68.98% ±0.35%	0.69 ±0.00	0.22 ±0.00
(b)	First 12 hours after admission	68.64% ±0.48%	24.78% ±0.27%	68.83% ±0.56%	68.63% ±0.53%	0.69 ±0.00	0.21 ±0.00
	First 24 hours after admission	68.91% ±0.33%	24.52% ±0.25%	70.44% ±0.55%	68.81% ±0.37%	0.70 ±0.00	0.21 ±0.00
	First 48 hours after admission	69.97% ±0.34%	23.81% ±0.16%	67.78% ±0.53%	70.11% ±0.40%	0.69 ±0.00	0.21 ±0.00

Table 6: Average confusion matrices across 10 runs for Experiment II with feature set (b) and using the data from the first 12 hours after admission. Average number of false negatives/positives and true negative/positives are rounded to whole numbers.

		Predicted+	**Predicted-**	
Balanced training set	**Condition +**	317	143	**Accuracy** ≈ 69% **F1 score** ≈ 25% **Sensitivity** ≈ 69% **Specificity** ≈ 69% **AUC** ≈ 0.69 **MCC** ≈ 0.21
	Condition -	1778	3893	
Unbalanced training set	**Condition +**	9	451	**Accuracy** ≈ 93% **F1 score** ≈ 5% **Sensitivity** ≈ 2% **Specificity** ≈ 100% **AUC** ≈ 0.51 **MCC** ≈ 0.10
	Condition -	5	5666	

Table 7: Mean performance and 95% CIs across 10 runs for Experiment III with feature set (b).

Data Collection Window	**Accuracy**	**F1 score**	**Sensitivity**	**Specificity**	**AUC**	**MCC**
12 hours leading to sepsis	98.63% ±0.17%	98.74% ±0.15%	99.74% ±0.13%	97.23% ±0.26%	0.98 ±0.00	0.97 ±0.00
24 hours leading to sepsis	97.07% ±0.26%	97.19% ±0.27%	99.76% ±0.14%	94.30% ±0.47%	0.97 ±0.00	0.94 ±0.00
48 hours leading to sepsis	92.40% ±0.44%	92.38% ±0.60%	97.84% ±0.31%	87.52% ±0.67%	0.93 ±0.00	0.85 ±0.01

Table 8: Mean performance and 95% CIs across 10 runs for Experiment I with feature subset (b) and using the data from the first 12 hours after admission, developed and tested cross hospital sizes.

Training on patients admitted to	**Total number of hospitals**	**Model accuracy when tested on patients admitted to**		
		Small hospital	**Medium hospital**	**Large hospital**
Small hospital	292	0.6220±0.011	0.6505±0.003	0.6333±0.005
Medium hospital	68	0.6202±0.004	0.6416±0.015	0.6475±0.004
Large hospital	17	0.6074±0.005	0.6392±0.004	0.6743±0.019

Table 9: Top ten most important contributing factors to sepsis/mortality prediction for a subset of random forest models. The numbers in the table indicate the rank of the features (if among top 10) and their average importance across 10 runs for the corresponding experiment.

Features	**Experiment I with feature subset (b) using the first 12 hours after admission**	**Experiment II with feature subset (b) using the first 12 hours after admission**	**Experiment III with feature subset (b) using the 12 hours leading to sepsis**
Entropy of respiratory rate	1 (0.0799)	1 (0.0838)	-
Mean of heart rate	2 (0.0725)	9 (0.0484)	7 (0.0596)
Maximum of heart rate	3 (0.0668)	-	1 (0.1855)
Entropy of temperature	4 (0.0585)	8 (0.0502)	-
Minimum of WBC count	5 (0.0535)	-	6 (0.0729)
Standard deviation of temperature	6 (0.0533)	-	9 (0.0377)
Maximum of WBC count	7 (0.0527)	10 (0.0445)	3 (0.0958)
Mean of WBC count	8 (0.0519)	-	4 (0.0811)
Mean of temperature	9 (0.0514)	4 (0.0617)	-
Entropy of heart rate	10 (0.0501)	3 (0.0618)	-

Mean of respiratory rate	-	2 (0.0660)	-
Standard deviation of respiratory rate	-	5 (0.0588)	8 (0.0457)
Minimum of temperature	-	6 (0.0515)	5 (0.0744)
Maximum of respiratory rate	-	7 (0.0503)	2 (0.1552)
Maximum of temperature	-	-	10 (0.0322)