Surviving Sepsis – Better Outcomes Through Early Detection with Machine Learning

Rhia E. Morrissette

Western Governors University

**Table of Contents**

[Project Overview 3](#_Toc77701925)

[A. Project Highlights 3](#_Toc77701926)

[Project Plan 3](#_Toc77701927)

[B. Project Execution 3](#_Toc77701928)

[Methodology 4](#_Toc77701929)

[C. Data Collection Process 4](#_Toc77701930)

[C1. Advantages and Limitations of Data Set. 4](#_Toc77701931)

[D. Data Extraction and Preparation Processes 4](#_Toc77701932)

[E. Data Analysis Process 7](#_Toc77701933)

[E1. Data Analysis Methods 7](#_Toc77701934)

[E2. Advantages and Limitations of Tools/Techniques 7](#_Toc77701935)

[E3. Application of Analytical Methods 7](#_Toc77701936)

[Results 8](#_Toc77701937)

[F. Project Success 8](#_Toc77701938)

[F1. Statistical Significance 8](#_Toc77701939)

[F2. Practical Significance 10](#_Toc77701940)

[F3. Overall Success 10](#_Toc77701941)

[G. Key Takeaways 10](#_Toc77701942)

[G1. Summary of Conclusions 10](#_Toc77701943)

[G2. Effective Storytelling 11](#_Toc77701944)

[G3. Findings-based Recommendations 11](#_Toc77701945)

[H. Panopto Presentation 11](#_Toc77701946)

[Appendices 11](#_Toc77701947)

[I. Evidence of Completion 11](#_Toc77701948)

[Sources 11](#_Toc77701949)

# Project Overview

# Project Highlights

* Research Question

Create a machine learning algorithm that can identify patients at high risk for sepsis based on signs and symptoms earlier than clinicians would. This enables earlier treatment and better chances for survival from sepsis.

* Project Scope
  + This project includes measurement of clinician time to identification as well as accuracy of identification. Because of multiple laws around patient data and confidentiality we will not be able to utilize real patient data for this project. For purposes of defining “clinician” identification accuracy we will use data from the work done by Goh, et al (2021) that reviewed about 300 sepsis cases and evaluated the accuracy of provider identification. The project includes development of a machine learning algorithm to identify sepsis. Again, due to HIPAA regulations and other privacy laws, we will not use actual clinical data. We will use this publicly available data set from Kaggle that includes de-identified patient data. Development of the algorithm will include creating a training and testing set, feature analysis and identification of the best parameters and cross validation of the algorithm.
* Solution Overview - Tools and Methodologies
  + There is a clear need for timely intervention in patients identified with sepsis. In a busy healthcare environment with system automation (vital signs, lab results, etc.) clinicians may not be immediately aware of sepsis onset. An automated system notification based on a machine learning algorithm can alert clinicians in a faster, more efficient manner to prioritize patient assessment and treatment.
  + For this project, we will use the waterfall methodology since this work fits well with the linear, sequential timeline associated with it. Often in illustrations, the five phases of this methodology, Requirement Gathering, Design, Implementation, Verification and Maintenance are depicted in a descending cascade, hence the methodology naming.

# Project Plan

# Project Execution

Summarize how the execution of each of these elements differed from the plan presented in Task 2.

* Project Plan
  + Overall, the project plan presented in Task 2 was followed although with longer time frames for some tasks.
* Project Planning Methodology
  + Our methodology, Waterfall, worked very well here. The data was collected, cleaned and analyzed. The analysis was validated and visualizations created. There could be some space for the Agile methodology in the development process. When fine tuning algorithm and PCA, there were several iterations of updating test data, running the algorithm and assessing it that would fit well with that methodology.
* Project Timeline and milestones
  + It took a bit longer to actually clean the data and code the algorithms and validation than expected. However, the other estimates were pretty accurate. All together the difference added 8 days to the project timeline. All milestones were accurate and completed.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Milestone** | **Projected Start Date** | **Projected**  **End Date** | **Planned Duration** | **Actual Duration** | **Difference** |
| Design Algorithm | 6/21/21 | 6/23/21 | 3 days | 2 days | -1 Day |
| Build Algorithm | 6/23/21 | 6/25/21 | 2 days | 10 days | +8 days |
| Clean / Format Data | 6/25/21 | 6/26/21 | 1 day | 2 days | +1 day |
| Train Algorithm / Cross Validation | 6/26/21 | 6/26/21 | 1 day | 1 day | 0 days |
| Analyze results | Visualization | 6/26/21 | 6/28/21 | 2 days | 2 days | 0 days |

# Methodology

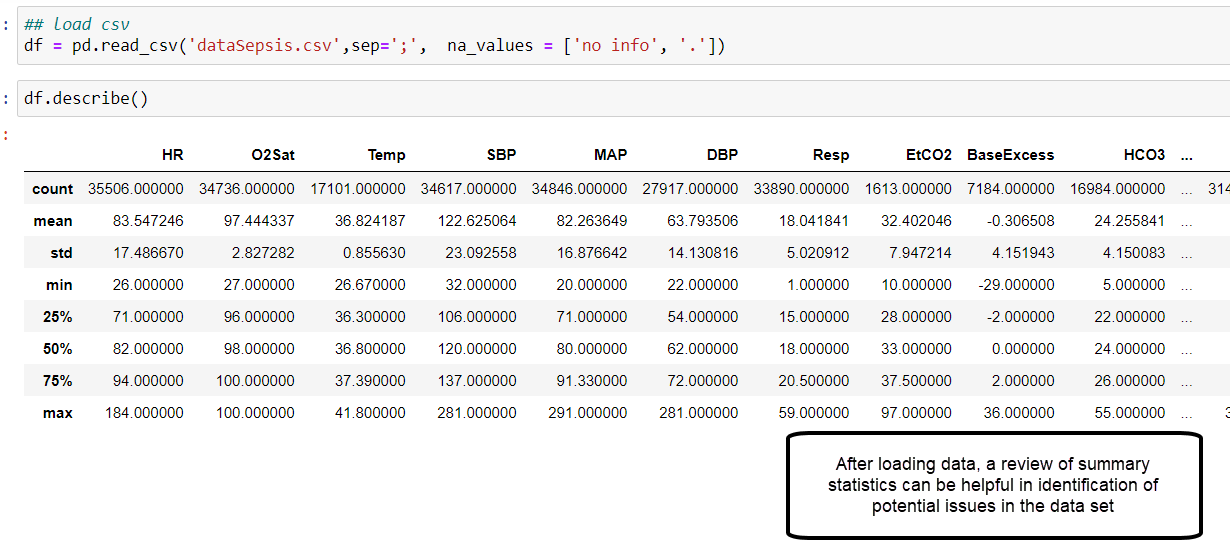
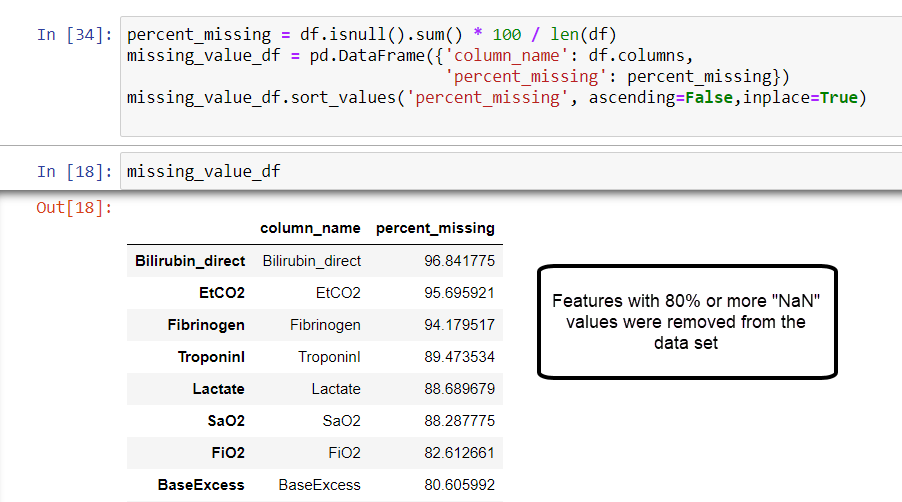
# Data Collection Process

* The data selection process for this project had no differences from the planned collection process. The data set was a previously collected set of patient data. The data selection process for me was downloading the csv file from the Cardiology Computing Challenge.
* There were no obstacles to my data collection process.
* There were no unplanned data governance handling issues.

# C1. Advantages and Limitations of Data Set.

* When reviewed, this data set presented some advantages but also some limitations for the analysis being performed. There were plenty of elements that could be used as features for the machine learning algorithm – 40 in all. This is a good number of features as it gives us flexibility in determining the best features to use for a classification algorithm. However, as often is the case, the good is also the bad. 40 features are way too many for an accurate algorithm and will need extensive review and feature engineering.

# Data Extraction and Preparation Processes

* Once the data set was downloaded in a CSV format, I imported it in to a python dataframe using the Pandas library. This open-source library was specifically designed to facilitate data analysis and has many great built-in tools and functions that to expedite data preparation so was very appropriate for the task. The key to good data analysis and machine learning especially relies on clean data.
* First, I did some overall review of the data set to get a general feel for the data. This is always a good idea and some basic summary statistics can quickly highlight areas of concern in the data.
* The first issue noted in this review was a large number of fields without values. There were 36302 records in the data set and some columns had more than 33,000 “NAN” (Not a Number) values. With very little information to provide our algorithm, these fields were of little worth to include. How to make the decision though? After thought and review of some Machine Learning guides, I decided to remove any columns/features that were 80% or more “NAN” values. Unfortunately, this list included EtCO2 and Lactate values which are some of the key indicators of sepsis in a patient. However, as the initial competition noted, these values were left blank as from their initial source purposefully and it make sense. If sepsis isn’t suspected by the human caregiver (as is our hypothesis) in a timely manner, these labs likely wouldn’t be ordered on a patient.
* After limiting the data set, further review of our summary statistics also revealed some further anomalies that required cleaning.
  + The minimum values for several of the vital statistics were so low as to be physically incompatible with life.
    - The minimum HR in the data set was 26 bpm (Normal 60-100 for adults). Bradycardia or a slow heart rate can occur but usually by the time it’s in the 40-bpm range, patients are symptomatic.
    - The minimum respiratory rate was 1 br/min; again, this is far below the normal adult range of 16-20 br/min. The data set doesn’t include information regarding use of mechanical ventilation but there are clinical instances where patients are not breathing on their own at all. In these instances, ventilators are typically set to breathe 10-12 times per minute for the patient.
    - SBP (Systolic Blood Pressure), DBP (Diastolic Blood Pressure) and MAP (Mean Arterial Pressure) and Oxygen Saturation had similar issues with the minimum value in each case being so low as to be unlikely.
  + The question became what to do with the records with the anomalous readings. Ideally, we would discard them since they would provide potentially erroneous data to our algorithm and thus decrease efficacy. However, if we remove too many records from the data set, we also run the risk of not having a good fit.
  + After reviewing the affected records manually, I discovered
    - Many of the records that had erroneous data for one field also had erroneous data in one or more of the other fields so there was considerable overlap.
    - All in all, to remove these records would remove less than 2000 records from the data set leaving us with over 34,000 records to use from training/testing. This is an adequate amount for a classification algorithm
  + Criteria used to remove rows was based on my own clinical knowledge and background as a nurse of 30 years with almost 10 years in ICU nursing.
    - HR less than 40, Respiratory Rate less than 10, SPB less than 80, DPB less than 40 and Oxygen Saturation less than 75% were removed. I didn’t remove rows based solely on the MAP as this is a calculated value based on the SBP and DBP measures so removing the erroneous values there should mostly have compensated for the MAP variances seen.
* Finally, there were several features included in the data set that while of use in the original contest using it, are not needed here. These include time of admission and unit. Originally the date set was created for real time sepsis identification and each data point had a person and timestamp indicator. Time of admission to a specific unit was used to determine the human time sepsis was identified. We are not performing real time analysis in this project – just simple detection. Therefore, the features were superfluous and discarded.

# Data Analysis Process

# E1. Data Analysis Methods

For this analysis, we utilized supervised learning methods for classification. Specifically, we used Support Vector Machines, Gaussian Naïve Byes and Decision Trees to see which provides the best ratings when evaluated. These supervised learning algorithms were appropriate for this analysis because we provided inputs (in the form of patient data) and also gave the algorithm the desired output – the criteria for a diagnosis of sepsis. This was superior to an Unsupervised learning since we have a specific classification we’re looking for in this analysis.

# E2. Advantages and Limitations of Tools/Techniques

There are many advantages and few limitations in using the supervised learning methods listed above. Advantages include availability of resources, ease of use and evaluation, standardization and scalability. Data analysis in general and machine learning specifically have grown in use and importance in this era of “Big Data” and there are many resources available to identify use cases and help interpret data. Machine Learning in particular has the benefit of the SKLearn library used in this project. This open-source Python library was created for machine learning algorithms and their evaluation. It has many easy-to-use functions and build across different algorithms are similar, increasing ease of use. They also have built in, standardized, statistical analysis tools to evaluate algorithm performance so comparisons of effectiveness are able to be done easily and efficiently. Finally, the algorithms are scalable and can be used on small or large datasets with little impact to performance- we look at scalability later in this report. The main limitation of these algorithms is their simplicity. While they are adequate for the work being done here, simple classification based on a static set of data points, they would not be adequate for the same classification in real time using real time patient data. The winners of the 2019 competition developed a new signature-based regression model to adequately evaluate disease progression in real time (Morrill, et al, 2019).

# E3. Application of Analytical Methods

The first step in application of supervised learning algorithms after data wrangling is to put the data in to a format that works well with the SKLearn library. While a pandas dataframe is similar to an array, it’s ideal to have the data in a true array. Also, we need to separate the features from the indicator (sepsis y/n). To do this, I created 2 arrays, x2 includes all features previously discussed and y2 containing only the sepsis indicator.

Once the arrays were created, I then used SKLearn test\_train\_split feature to break this in to 4 arrays – X\_train, X\_test, y\_train, y\_test – before performing any further analysis. This step should be done prior to any further feature engineering to avoid unnational effects on the test set. In other words, we want to keep the test data set as pristine as possible for comparison purposes. We did this using a split of about 30% - 30% of our data was put in to the test set and 70% in the training set and we set a random seed of 42 for reproducibility.

I decided to run the algorithms on the data sets as is so we could see the impact of feature engineering. As previously stated, we ran 3 algorithms on the data, Gaussian NB, Decision Tree and SVM. All 3 algorithms showed an accuracy rate of 0.9 or greater which seems great on the face of it. However, accuracy provides an overall number of how often the algorithm is correct in its classification of all data. But we are more concerned with correctly identifying sepsis when present than correctly determining it’s not. So, for this analysis, recall (the number of true positives divided by true positives plus false negatives) is a more important identifier. All 3 algorithms produced a recall score greater than 0.6 with the Gaussian algorithm coming in at a high of 0.77. Already, our algorithms are outpacing the rate at which humans identify sepsis using patient characteristics which is set at a threshold of 0.5.

Now that we established an adequate accuracy and recall in our algorithm right from the gate, we checked to see if we had opportunities for improvement. For primary component analysis, we used Select Kbest a univariate feature selection tool that is part of the SKLearn library. It removes all but the K best features from the features array. To do this, first we find the features that have the highest f-score. Many scores can be used here – I used the f-score as this provides an ANOVA F value between label/feature in classifications which is on point here. Looking at the list provided here, I played with the numbers a bit and determined 22 would be a reasonable number of features to provide to keep good accuracy and recall. I used the SelectKBest functions to identify the 22 best (of 28) features to use for analysis and transformed my initial data sets to only include them. From here, the data was resplit in to train/test sets using the same parameters previously applied. Using these updated sets, I reran all 3 algorithms with minimal improvement in any of the statistics – less than 0.1 change in most cases. Gaussian NB is still our algorithm with the best ratings so decided to focus here moving forward.

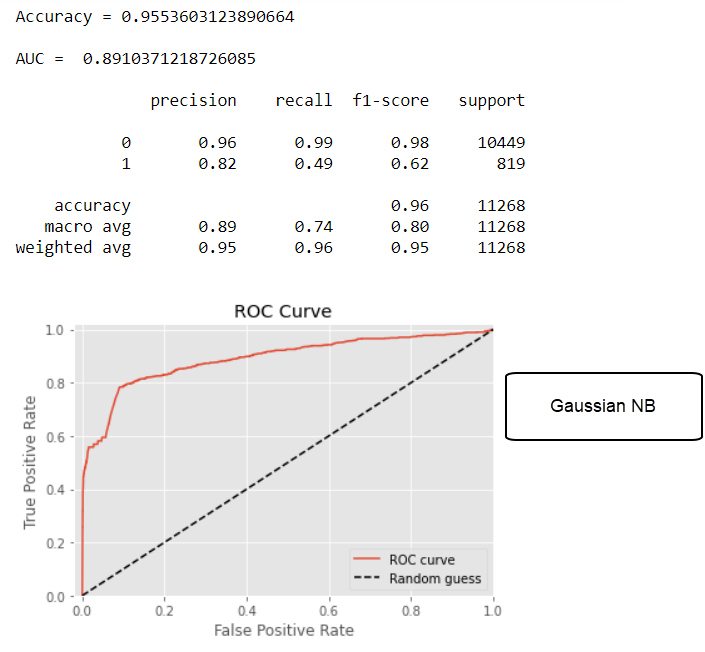
To validate and tweak the number of features I selected manually using Select Kbest, I created a pipeline and used GridSearchCV to test all possible “K” values against the GaussianNB algorithm. Surprisingly, this returned an optimal K value of 2 (from a possible 28) features. After rechecking programming, this did appear to be the recommendation from GridSearchCV. While feeling very counterintuitive, I updated the data sets using the 2 KBest features only, created train/test sets and ran the Gaussian MB algorithm again. There was improvement in our AUC score when compared to all features as well as the Kbest of 22 but it was minimal at best – 0.89 vs 0.88).

# Results

# Project Success

# F1. Statistical Significance

Multiple statistics were used here to evaluate the efficacy of modeling. Three metrics will be used to evaluate effectiveness of the algorithm, accuracy, precision and recall. Accuracy is the sum of true positives and true negatives divided by the total number of predictions and is one of the basic measurements of a machine learning algorithm. More simply, it is how many times did the algorithm correctly identify the classification for a data point provided. Precision measures how many items labeled as positive (part of the expected class) were truly positive and it is a calculation of true positives divided by the sum of true positives and false positives. We can see that this is an important measure when evaluating a classification algorithm as we’re creating here. Recall is another standard metric when evaluating machine learning algorithms and is a measure of how many of the positive items, the algorithm correctly identified. In other words, how many positive items did the algorithm recall from the data set. Recall is calculated as true positives divided by the sum of true positives and false negatives. As Rawat (2019) states, “While recall expresses the ability to find all relevant instances in a dataset, precision expresses the proportion of the data points our model says was relevant actually were relevant.”. A final metric to evaluate our algorithm will be the F1 score. This score is a mean of recall and precision so somewhat mitigates the inverse relationship between them. These are all appropriate, industry standard methods and metrics for evaluation of machine learning algorithms. Additionally, I looked at the AUC (Area Under Curve). This value looks at the area under a ROC (Receiver Operating Characteristic) Curve which compares the probability of our algorithm performing better than a guess. Since a guess between 2 items as a probably of 0.5, an AOC greater than 0.5 is better than guessing and the closer it is to 1, the better it is.

For each iteration of an algorithm, the same statistics were evaluated. An overall accuracy, precision and recall rate and F1 score. I also calculated the precision and recall for each class, “Not Sepsis” (0) and “Sepsis” (1) to evaluate correct identification of sepsis patients. I also calculated the ROC curve and the resulting AUC calculation. An example of the output from the calculations performed using the SKLearn library is shown here.

All of the algorithms had AUC scores well above the 0.5 threshold for and fairly close to each other as shown in the graphic below (the scale is a difference of 0.02 between the top 2). Decision Tree overall performed poorly in this comparison and may be too simplistic in general for this type of classification.

In the final analysis, we have significantly better precision and recall than the human analysis in detecting sepsis.

# F2. Practical Significance

Practically speaking, the significance of this work shows that even simple machine algorithms can improve detection of sepsis and this translates to live saved. Septic shock, a subset of sepsis that has “profound circulatory, cellular, and metabolic abnormalities” (Singer et al., 2016), cause more than 250,000 deaths in the US each year according to the CDC. As the Surviving Sepsis guidelines point out, early detection and treatment are key in preventing mortality. If a machine algorithm can detect sepsis even 1% better than today’s clinicians that translates to 2500 lives saved.

# F3. Overall Success

Overall, the project was successful as it disproved our null hypothesis that a Machine learning algorithm will perform no differently in identifying possible cases of sepsis than a human clinician. with statistical significance.

# Key Takeaways

# G1. Summary of Conclusions

Conclusion from this project

* Even in the simplest exercise machine learning can detect sepsis from signs and symptoms better than clinicians.
* This improved detection translates to lives saved through early clinical intervention
* The ease of creation and use of existing machine learning libraries for this work allow for scalability and use in almost any clinical setting – no super computer needed

# G2. Effective Storytelling

The tools used for this project make sense from a story telling perspective as they clearly show the statistical basis for the conclusions in a way easily interpreted by readers.

# G3. Findings-based Recommendations

This project is only the beginning of the work to improve sepsis detection and early treatment using machine learning. This project did a retrospective review – it looked at patients that had been identified as septic or not and learned to classify them based on the features provided. This is helpful in determining feasibility of using machine learning but needs to be taken to the next level. This would be a more in-depth analysis of the features presented here – determining why some of the more important clinical indicators (lactate, EtCO2) had such high percentages of null values and perhaps better account for that as well as identify if any additional features would benefit the training of the algorithm. Finally, translating this work to an algorithm capable of accepting and analyzing data in real time is key. More research should be done to determine how early a machine algorithm accurately identify not only if sepsis is present but if clinical values are trending *towards* sepsis. The ultimate goal – prevention.

# Panopto Presentation

Provide a link to your Panopto presentation. Include the following in your summary:

[Morrissette Panopto Presentation](https://wgu.hosted.panopto.com/Panopto/Pages/Sessions/List.aspx?folderID=b9e7c777-c0e0-4dc6-90dd-ac68013a3cc2)

# Appendices

# Evidence of Completion

Python Code

CSV of Data Used

Screenshots of Output used throughout document and presentation

# Sources

Morrill, James & Kormilitzin, Andrey & Nevado-Holgado, Alejo & Swaminathan, Sumanth & Howison, Sam & Lyons, Terry. (2019). The Signature-Based Model for Early Detection of Sepsis from Electronic Health Records in the Intensive Care Unit. 10.22489/CinC.2019.014.

William P. T. M. van Doorn,Patricia M. Stassen,Hella F. Borggreve,Maaike J. Schalkwijk,Judith Stoffers,Otto Bekers,Steven J. R. Meex. A comparison of machine learning models versus clinical evaluation for mortality prediction in patients with sepsis. PLOS: January 19, 2021. <https://doi.org/10.1371/journal.pone.0245157>

McCoy, A. D. (2017). Reducing patient mortality, length of stay and readmissions through machine learning-based sepsis prediction in the emergency department, intensive care unit and hospital floor units. *BMJ, 6*(e000158), . doi:10.1136/bmjoq-2017-000158

Vincent J-L (2016) The Clinical Challenge of Sepsis Identification and Monitoring. PLoS Med 13(5): e1002022. <https://doi.org/10.1371/journal.pmed.1002022>

Singer M, Deutschman CS, Seymour CW, et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA.* 2016;315(8):801–810. doi:10.1001/jama.2016.0287

Bai, X., Yu, W., Ji, W. *et al.* Early versus delayed administration of norepinephrine in patients with septic shock. *Crit Care* **18,**532 (2014). <https://doi.org/10.1186/s13054-014-0532-y>

Dellinger, R. Phillip MD1; Levy, Mitchell M. MD2; Rhodes, Andrew MB BS3; Annane, Djillali MD4; Gerlach, Herwig MD, PhD5; Opal, Steven M. MD6; Sevransky, Jonathan E. MD7; Sprung, Charles L. MD8; Douglas, Ivor S. MD9; Jaeschke, Roman MD10; Osborn, Tiffany M. MD, MPH11; Nunnally, Mark E. MD12; Townsend, Sean R. MD13; Reinhart, Konrad MD14; Kleinpell, Ruth M. PhD, RN-CS15; Angus, Derek C. MD, MPH16; Deutschman, Clifford S. MD, MS17; Machado, Flavia R. MD, PhD18; Rubenfeld, Gordon D. MD19; Webb, Steven A. MB BS, PhD20; Beale, Richard J. MB BS21; Vincent, Jean-Louis MD, PhD22; Moreno, Rui MD, PhD23 and the Surviving Sepsis Campaign Guidelines Committee including the Pediatric Subgroup Surviving Sepsis Campaign, Critical Care Medicine: February 2013 - Volume 41 - Issue 2 - p 580-637

doi: 10.1097/CCM.0b013e31827e83af

Seymour CW, Gersten F, Prescott HC, et al. Time to treatment and mortality during mandated emergency care for sepsis. N Engl J Med 2017; 376(23):2235-2244

Machine Learning. (2021, June 20) In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Machine\_learning&oldid=1029354622

Goh, K.H., Wang, L., Yeow, A.Y.K. *et al.* Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. *Nat Commun* **12,**711 (2021). <https://doi.org/10.1038/s41467-021-20910-4>

Reyna MA, Josef CS, Jeter R, Shashikumar SP, Westover MB, Nemati S, Clifford GD, Sharma A. Early Prediction of Sepsis From Clinical Data: The PhysioNet/Computing in Cardiology Challenge. Critical Care Medicine 48 2: 210-217 (2019). https://doi.org/10.1097/CCM.0000000000004145