**DATA 624 - DATATHON #6**

**GROUP 1 - HANNAH DIMMICK, MINKI LEE, RACHEL NGUYEN**

**PREDICTING DEVELOPMENT OF SEPSIS IN PREVIOUSLY-ADMITTED ICU PATIENTS**

**Introduction**:

Sepsis, an extreme and life-threatening response to infection, is an incredibly dangerous condition with a ~22% mortality rate worldwide (although mortality rates differ drastically depending on access to medical care) [1]. It is the most common cause of admission to an intensive care unit (ICU), and the most common cause of death in the ICU [2]. Additionally, the risk of developing sepsis while staying in the ICU for any reason is incredibly high; in one study analysing 170 post-surgical patients who were treated in the ICU, 83 (49%) developed sepsis within the 28-day monitoring period [3]. There are also significant long-term side effects associated with sepsis development even if the patient recovers and is discharged from the ICU [4,5,6]. While the markers of mortality risk post-sepsis development are relatively well-known [7], the bigger challenge and more clinically relevant problem is to determine how to monitor patients *prior* to sepsis development. The importance and difficulty of accurate monitoring for this issue are significant factors in why Grand View Research, Inc. has estimated that the global sepsis diagnostics market is expected to be worth $USD 1.18 billion by 2027 [8].

The main drivers for this increase in market value are predicted to be blood culture media and assay kits/reagents [8]. While these specific tools would certainly be a worthwhile investment for many hospitals, any effort to determine accurate diagnostic indicators based on tools and tests that are already available and widely used would be extremely helpful. This is true for clinicians in more remote locations, but also for patients without access to health insurance who would have to pay for these additional, potentially pricey, tests out of pocket.

Only a small percentage of the projected market growth for sepsis diagnosis is expected to be in software [8]. However, we believe there are potential machine learning applications to the determination of risk of sepsis while in the ICU. To try to solve this problem, we looked at demographic and physiologic data from ICU patients that did or did not develop sepsis during their stay in order to determine if there were any key variables that could be used for monitoring early warning signs of sepsis, and hopefully prevent its occurrence.

**Data Engineering:**

The dataset used was “Sepsis-Data-Raw-A.” Although there were two datasets available (A and B), only one dataset was selected due to the large file size of each and the significant amount of processing time that would have been required to analyse the entirety.

To prepare the data for the random forest classifiers, the patient’s ID (patient), ICU length of stay (ICULOS), heart rate (HR), Pulse oximetry (O2Sat), systolic blood pressure (SBP), mean arterial pressure (MAP), Respiration rate (Resp), age, gender, and sepsis label were imported from the main dataset to Python. The patient, age, gender, and sepsis label categories were static for each patient, but the other variables were selected because they contained <20% missing data. The simple imputer was used to impute the mean into the non-static variables. Next, patients were excluded if they entered the ICU with sepsis, as these patients were not useful for a prediction task. From the original 20,336 patients in the dataset, 203 were excluded for arriving in the ICU with sepsis, leaving 20,133 unique patients for analysis.

Seven additional datasets were derived from the patient’s information at the first hour of admission to the ICU and at 1/2/3/4/5/6 hours prior to developing sepsis or being released from the ICU. In these datasets, each patient was represented with one line of data including all of the listed variables. Those who developed sepsis at any point were assigned a value “1” for the SepsisLabel column. Patients in each dataset that developed sepsis in a shorter time than the relevant hour range were excluded.

**Analysis:**

Data from the 1-hour prior to sepsis/release dataset was used to determine the descriptive statistics; the decision to exclude patients who entered the ICU with sepsis from the descriptive statistics was made for the same reason as for the random forest analysis: since the goal of this task was to predict development of sepsis in the ICU, patients who developed sepsis externally were not relevant to the question. Tableau Prep Builder was used to observe the distribution of the data and select information to create an infographic using descriptive statistics, which includes a bar chart, pie chart, and histogram about patients who were later diagnosed with sepsis.

Random forest models were created for each hour(s)-prior dataset individually. This was performed in Python using the RandomForestClassifier function from sklearn.ensemble. Features input were ICULOS, HR, O2Sat, SBP, MAP, Resp, age, and gender, with SepsisLabel being the target variable. Train-test split was performed using 0.75-0.25. Ten estimators were used, and the class weight was set as “balanced.” Dataframes and horizontal bar charts were generated to summarize the rank order of feature importance and the relative weight of each feature in the model. This process was performed again without including ICULOS. Several performance metrics were calculated for each model: accuracy, precision, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV).

Data at the ICU admission contained only one record for each patient where the ICULOS value is smallest. A random forest classifier was used to predict whether a patient would develop sepsis later using the admission data with imputed null values, and a Naive Bayes classifier was used for data with all null values dropped. Metrics of performance (ROC AUC, accuracy, precision, and confusion matrix) were evaluated.

**Findings:**

Of the patients who did not arrive in the ICU with sepsis, 1587 (7.88%) developed sepsis at least one hour into their stay, with 18,546 (92.12%) being released without ever developing sepsis. Among patients who developed sepsis, we found that sepsis occurred more frequently in male patients (61% male and 39% female). Older adults 60+ accounted for 61%, and age between 70-80 was the highest. Their average ICU length of stay was about 2.3 days (approx. 56.3 hours). Most patients were diagnosed with sepsis within a day, but some patients stayed in ICU for as long as two weeks. We observed a similar trend of heart rates between patients with sepsis and without sepsis; however, a very rapid heart rate of higher than 110 beats per minute was more common in patients with sepsis when compared with patients without sepsis. Furthermore, a higher percentage of abnormal respiration rate was also observed. Interestingly, oxygen saturation level, mean arterial pressure, and systolic blood pressure showed a similar trend between patients with and without sepsis.

The performance evaluation of the admission data classification models suggests that fitting non-imputed data increased the predicting ability (from 5-12% for each metrics). However, the confusion matrices show that these models could only predict 8-11% sepsis cases. Looking at the admission data, we found that the patterns of the most important features observed in patients with and without sepsis were not different, which may explain the poor predicting ability.

After exclusions, the number (%) of sepsis patients in the 2/3/4/5/6 hour(s)-prior datasets were 1540 (7.67%), 1494 (7.46%), 1463 (7.31%), 1426 (7.14%), and 1384 (6.94%), respectively.

Because 233 patients developed sepsis within ≤6 hours of admission to the ICU, these patients are represented in one of the hour(s)-prior datasets in addition to the first hour dataset.

The length of stay in the ICU was by far the greatest determining factor for risk of sepsis development. However, because the variability in this metric is extremely high (mean±st.dev.: non-sepsis = 37.9±13.9 hrs; sepsis = 56.3±54.4 hrs), and it is also not clinically beneficial knowledge for accurate prediction of sepsis, it was excluded and other factors were analysed independently of ICU stay length. There were no meaningful differences in the relative importance of each feature across different hour(s)-prior datasets. Age was consistently the greatest predictor of sepsis development (relative importance = 0.198-0.206), followed by heart rate (relative importance = 0.187-0.193). Gender was consistently the least important feature (relative importance = 0.022-0.026). The model with the highest accuracy and precision was 4 hours prior to sepsis/release (accuracy = 0.9704, precision = 0.9702). Sensitivity, specificity, PPV, and NPV were calculated and are summarized in Table 1.

| # Hour(s) Prior | Sensitivity | Specificity | PPV | NPV |
| --- | --- | --- | --- | --- |
| 1 | 74.8858447 | 98.8903394 | 86.5435356 | 97.6369495 |
| 2 | 78.7061995 | 98.4089443 | 79.7814208 | 98.3032646 |
| 3 | 72.752809 | 98.4959175 | 78.7234043 | 97.9277932 |
| 4 | 79.1891892 | 98.4675156 | 80.4945055 | 98.3401595 |
| 5 | 82.4512535 | 97.7772982 | 74.1854637 | 98.6286461 |
| 6 | 77.2079772 | 98.3160622 | 77.6504298 | 98.2736297 |

*Table 1: Sensitivity, specificity, PPV, and NPV for random forest models applied to hour(s) prior datasets*

**Limitations:**

Due to the fact of missing data in several of our variables of interest, imputing methods were used in order to create the random forest models. As imputing was performed at the first step in the data processing, we could not tell how many of the values imputed were in the hour(s) prior to sepsis/release datasets. Additionally, we do not know how many imputed values ended up being used for patients who developed sepsis vs. those who did not. To trust the conclusions of this analysis, we must assume that the imputed values were normally distributed across the entire dataset.

Another limitation in model creation was the fact that significantly fewer patients developed sepsis than did not. Adjustments in the model were made in order to account for this, but the result was a model that was still relatively low in sensitivity, likely due to this large imbalance. A possible alternative approach would have been to take a random sample of the non-sepsis patients that matched the number of patients in the sepsis group, and use that for comparison, but we chose to include as many patients as possible in this analysis.

**Conclusions:**

We looked at the number of patients who developed sepsis, and the gender proportion, age at risk, length of stay in ICU, heart rates, and respiration rate. We used machine learning models to determine the key variables that could be used for monitoring early warning signs of sepsis. The random forests were moderately successful at determining patients who would develop sepsis in the hour(s) leading up to development. However, although the specificity and NPV were excellent, the sensitivity and PPV were only moderate, which is problematic for the task at hand. In this situation, because of the danger associated with development of sepsis, sensitivity should be prioritized over specificity. Interestingly, it appears that this type of analysis can be done with similar levels of success 1-6 hours prior to sepsis development.

Age and heart rate were the most important indicators of potential sepsis risk in the hour(s) leading up to development, with MAP, SBP, and Resp being moderate indicators, and gender being a very unimportant indicator.

Giving appropriate treatment to patients with sepsis as soon as possible is important, but also the early prediction of sepsis is critical for treating life-threatening conditions as well. The early detection and diagnosis of sepsis can increase the likelihood of survival for patients with sepsis.Critical-care decision makers can use our findings as a reference for diagnosing sepsis ahead of time so that patients can receive the proper treatment at the right time.

**References:**

[1] *Sepsis Fact Sheet*. (2020, August 26). World Health Organization. https://www.who.int/news-room/fact-sheets/detail/sepsis

[2] Perner, A., Gordon, A. C., De Backer, D., Dimopoulos, G., Russell, J. A., Lipman, J., ... & Walsh, T. (2016). Sepsis: frontiers in diagnosis, resuscitation and antibiotic therapy. *Intensive care medicine*, *42*(12), 1958-1969.

[3] Pittet, D., Rangel-Frausto, S., Li, N., Tarara, D., Costigan, M., Rempe, L., ... & Wenzel, R. P. (1995). Systemic inflammatory response syndrome, sepsis, severe sepsis and septic shock: incidence, morbidities and outcomes in surgical ICU patients. *Intensive care medicine*, *21*(4), 302-309.

[4] Iwashyna, T. J., Ely, E. W., Smith, D. M., & Langa, K. M. (2010). Long-term cognitive impairment and functional disability among survivors of severe sepsis. *Jama*, *304*(16), 1787-1794.

[5] Winters, B. D., Eberlein, M., Leung, J., Needham, D. M., Pronovost, P. J., & Sevransky, J. E. (2010). Long-term mortality and quality of life in sepsis: a systematic review. *Critical care medicine*, *38*(5), 1276-1283.

[6] Sun, A., Netzer, G., Small, D. S., Hanish, A., Fuchs, B. D., Gaieski, D. F., & Mikkelsen, M. E. (2016). Association between index hospitalization and hospital readmission in sepsis survivors. *Critical care medicine*, *44*(3), 478-487.

[7] Pittet, D., Thievent, B., Wenzel, R. P., Li, N., Auckenthaler, R., & Suter, P. M. (1996). Bedside prediction of mortality from bacteremic sepsis. A dynamic analysis of ICU patients. *American journal of respiratory and critical care medicine*, *153*(2), 684-693.

[8] *Sepsis Diagnostics Market Size, Share & Trends Analysis Report By Product (Blood Culture Media, Software, Instruments, Assay Kits & Reagents), By Technology, By Pathogen, By Testing Type, And Segment Forecasts, 2021 - 2027*. (2021, February). Grand View Research. https://www.grandviewresearch.com/industry-analysis/sepsis-diagnostics-market?utm\_source=prnewswire&utm\_medium=referral&utm\_campaign=hc\_23-march-21&utm\_term=sepsis-diagnostics-market&utm\_content=rd/segmentation