Combining Process Mining and Machine Learning for Analysis of Adult Sepsis Events

Lauren Flemmer[1*], Hilda B. Klasky[2*]

[1]University of California, Riverside, Riverside, CA, 92521, USA
[2]Oak Ridge National Laboratory, Oak Ridge, TN 37830, USA
*flemmerlk@ornl.gov, klaskyhb@ornl.gov

INTRODUCTION

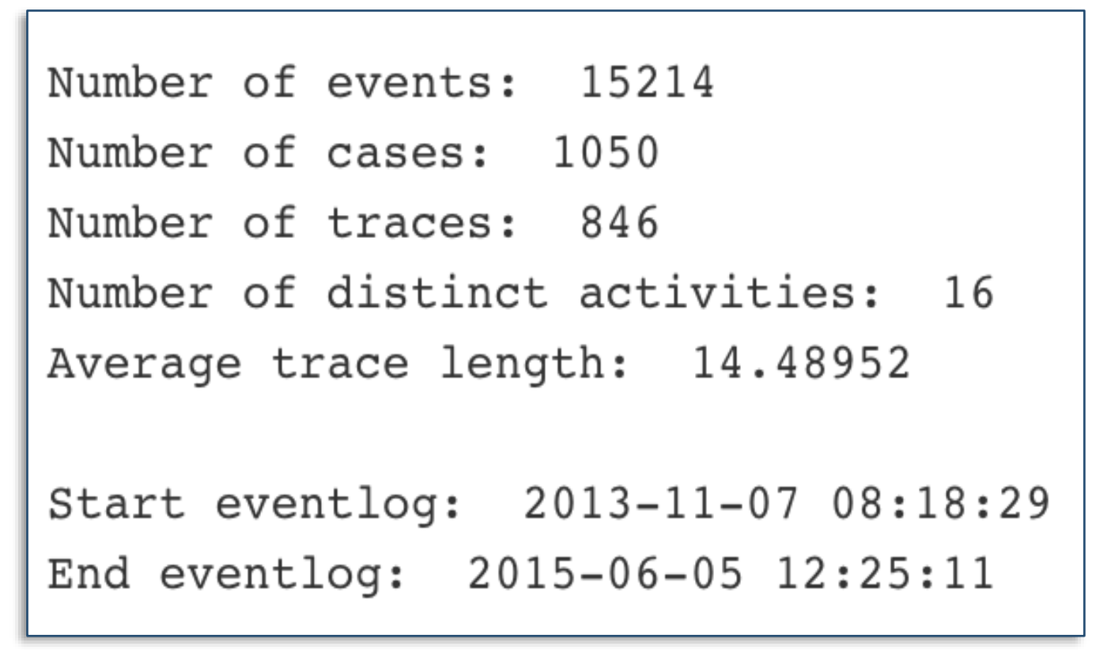
Process mining is a field of data science that allows us to **discover**, **visualize and analyze processes from event logs**. However, although process mining provides valuable and significant insights such as descriptive statistics, the process model map and where bottlenecks and re-work occurs in a process, it does not provide **probabilistic insights** about other features that influence the success and performance of the process.

RESEARCH OBJECTIVES

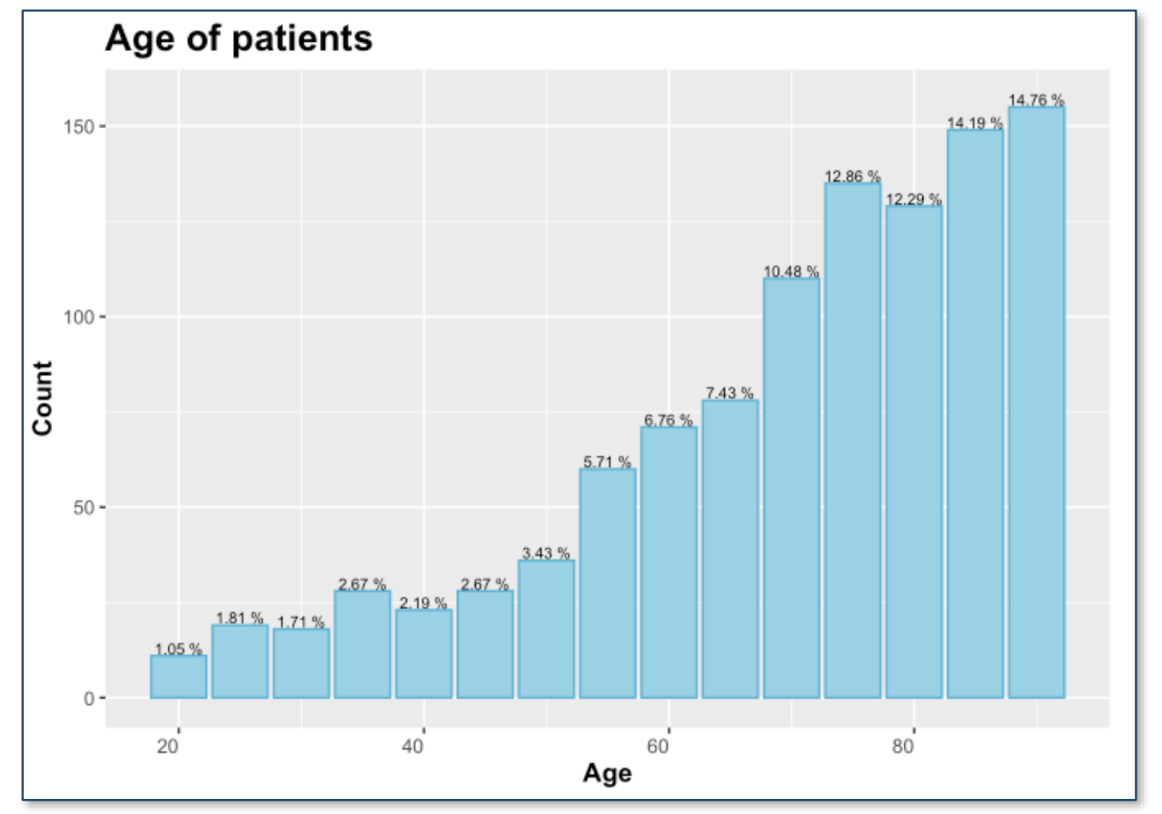
- To complement process mining, this study aims to explore, apply and combine process mining with machine learning to create statistical models that are both accurate and meaningful.
- The focus of that analysis was to better understand adult sepsis events as they flow in the care process and to develop machine learning models to predict patient discharge.

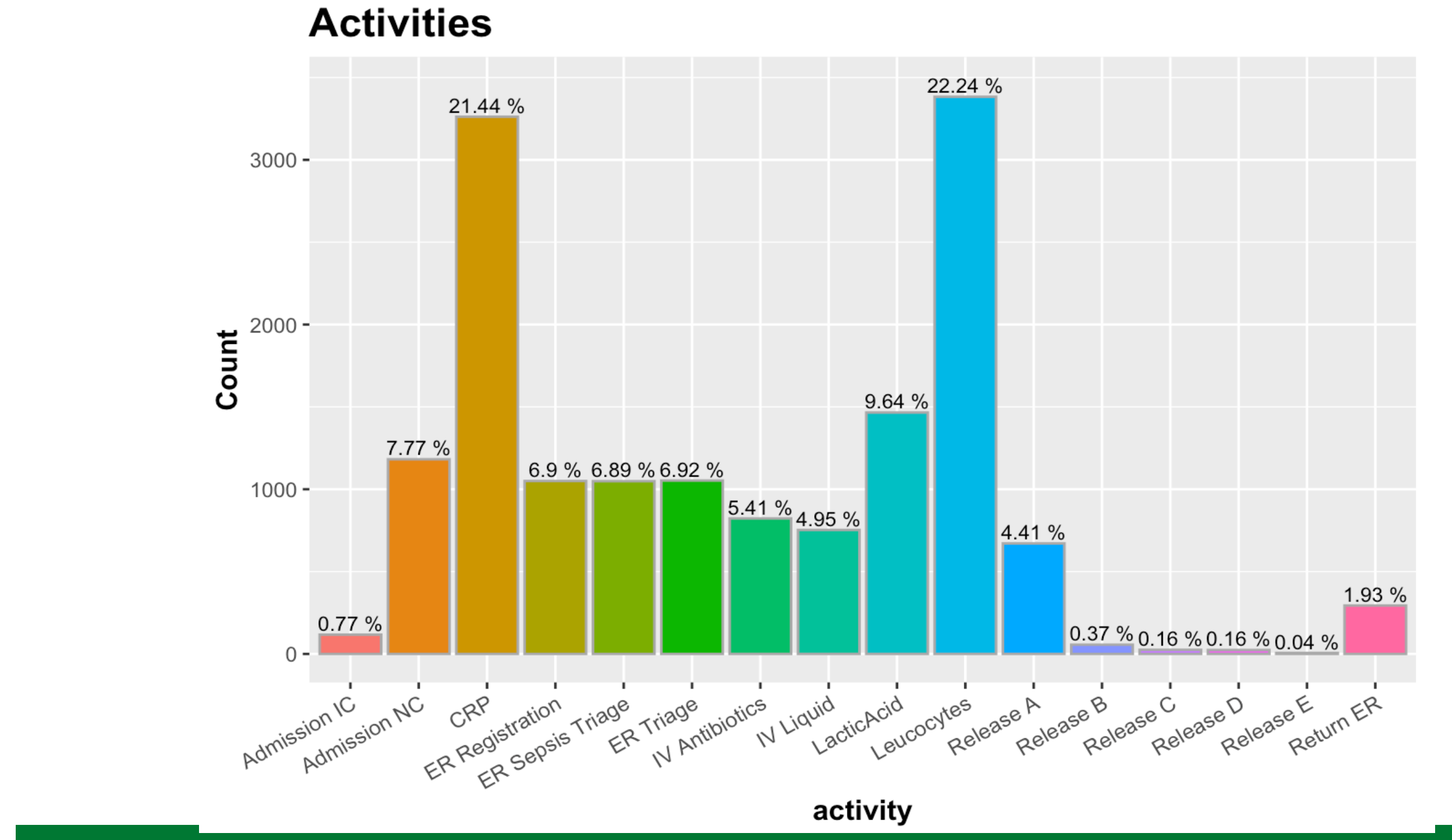
DATA

For a demonstration project, our initial efforts were directed to a dataset of adult sepsis events from health care records publicly available in the BupaR package.



```
Number of events:  15214
Number of cases:  1050
Number of traces:  846
Number of distinct activities:  16
Average trace length:  14.48952

Start eventlog:  2013-11-07 08:18:29
End eventlog:  2015-06-05 12:25:11
```



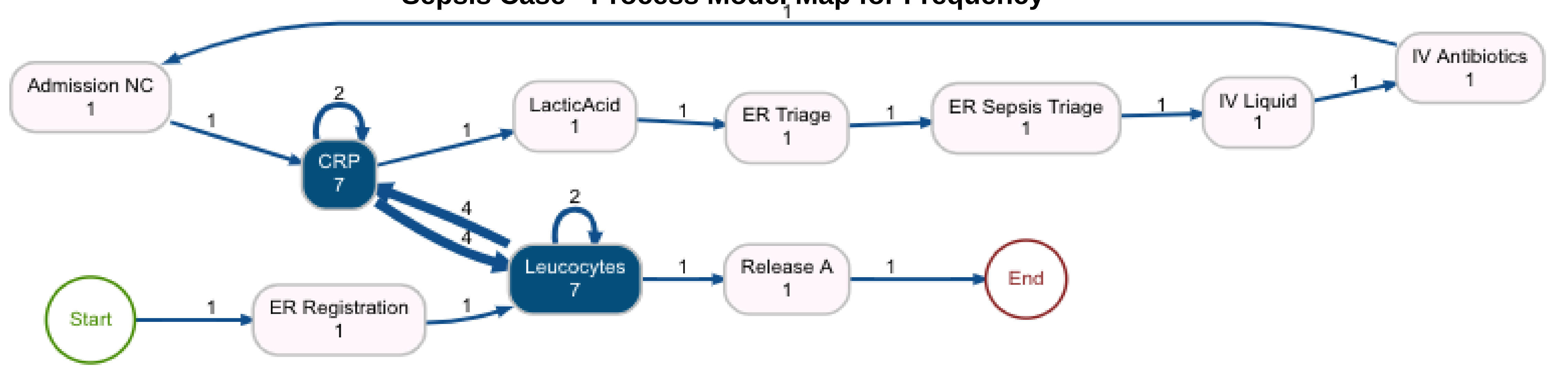


METHODS

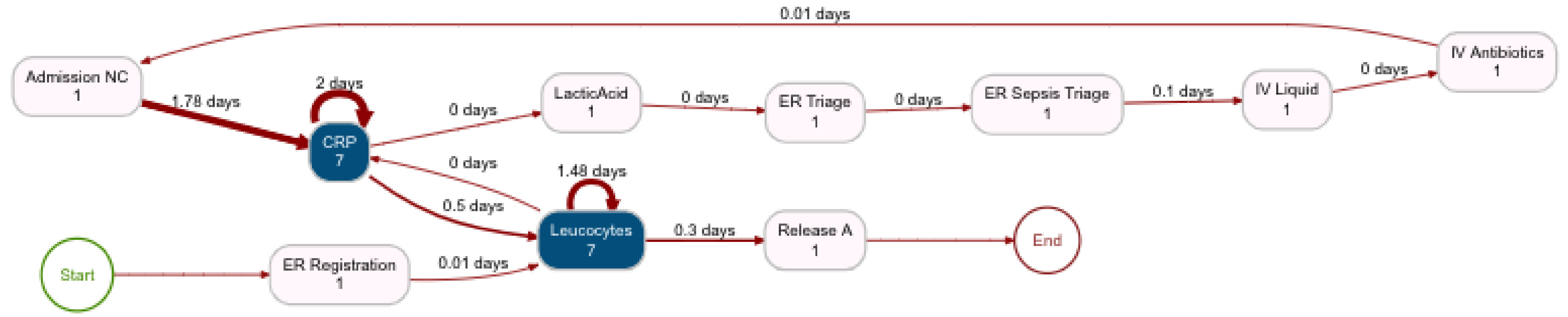
- **Process Mining:** Process maps illustrating the different types of hospital events were generated using the BupaR package to visualize the event log data and to gain an understanding of the overall flow of sepsis events, as well as the average time between each event/treatment.
- **Machine Learning:** We applied machine learning approaches, specifically logistic regression for classification and Markov chains for clustering, to analyze and visualize the dataset to obtain an understanding of its organization and the distributions of its variables and components. Finally, the models were trained and tested to accurately predict which hospital events are discharges, and which are not.

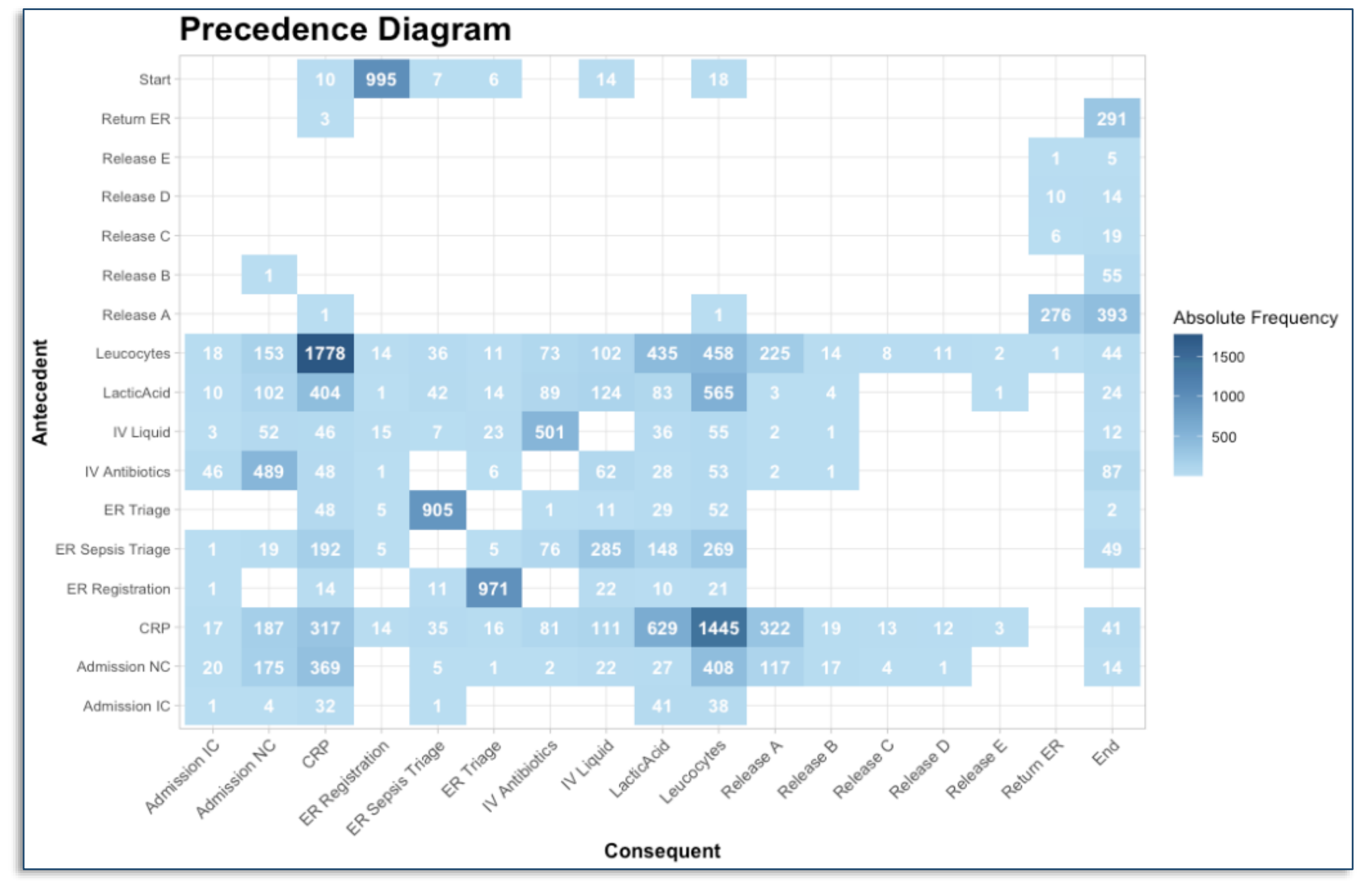
RESULTS I

Sepsis Case - Process Model Map for Frequency



Sepsis Case - Process Model Map for Performance

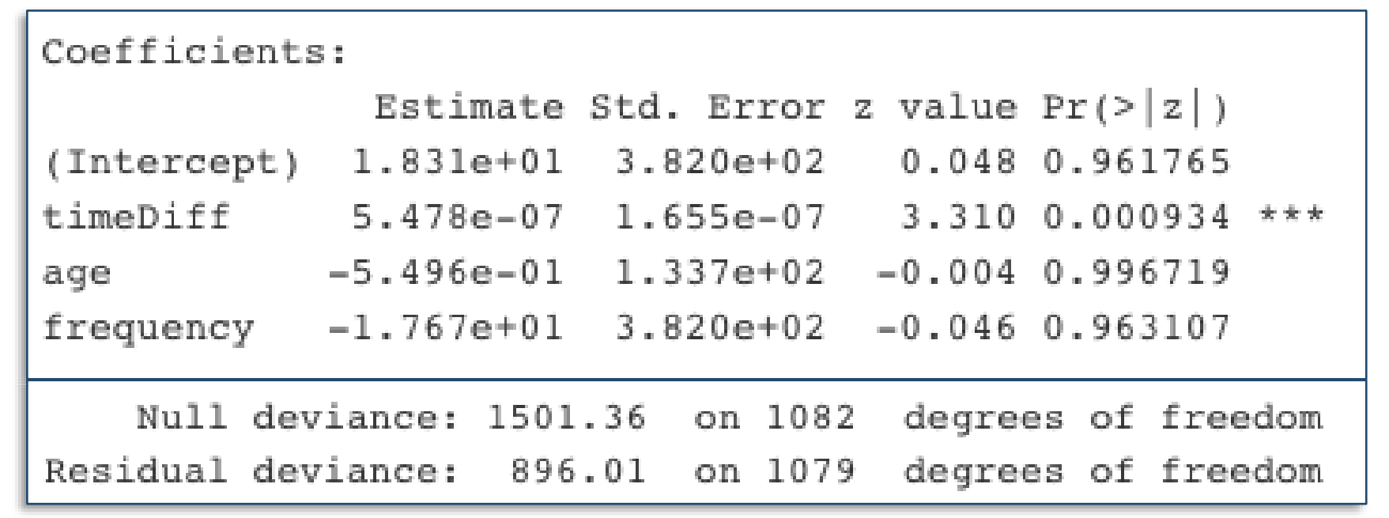




RESULTS II

Logistic Regression

- Classification accuracy refers to the % of events that were correctly predicted. In this case, **98.9%** of events were **correctly predicted** to be either **A Patient Discharge** or **Not a Patient Discharge.**



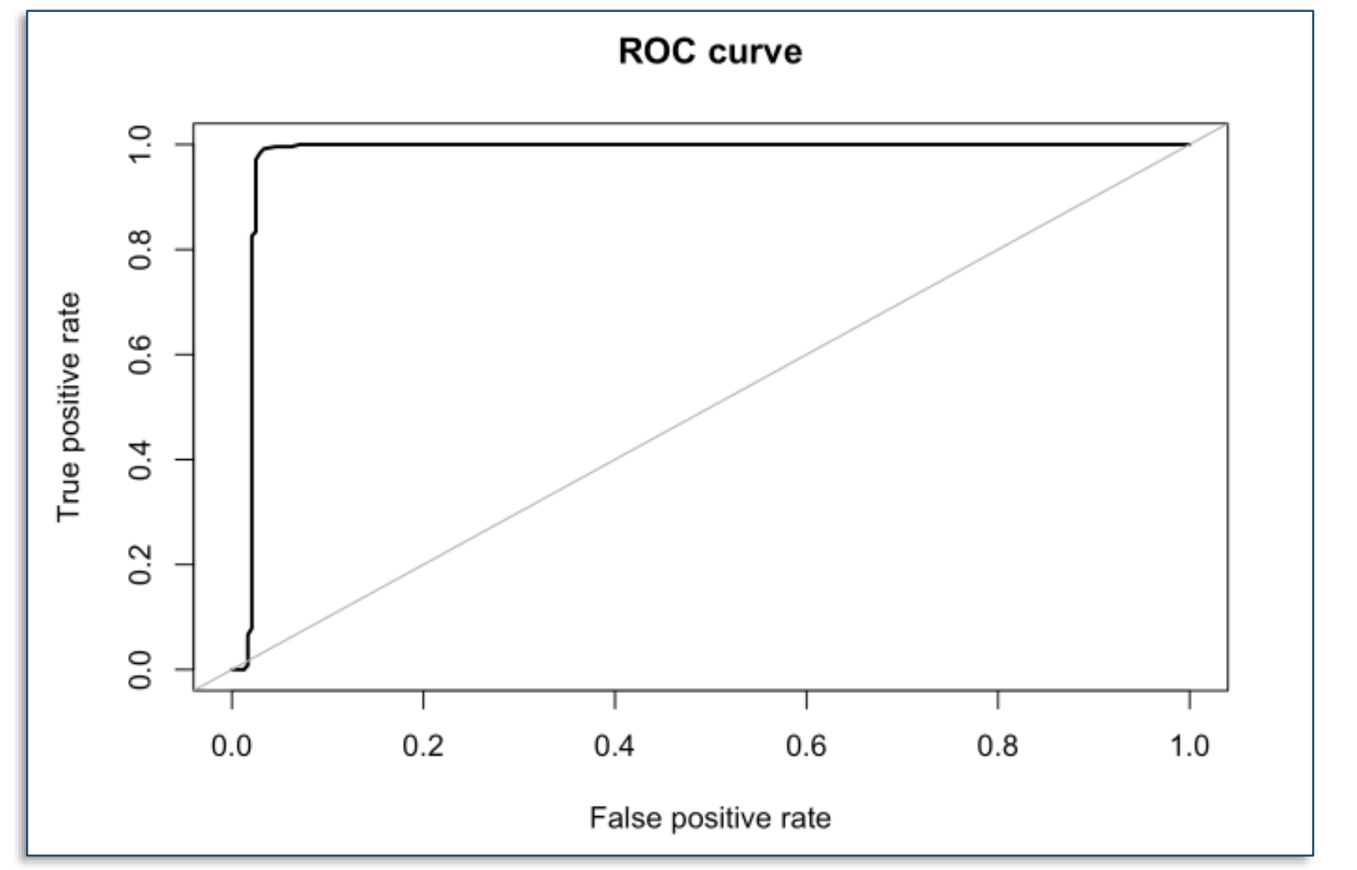
```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.831e+01  3.820e+02   0.048 0.961765
timeDiff     5.478e-07  1.655e-07   3.310 0.000934 ***
age         -5.496e-01  1.337e+02  -0.004 0.996719
frequency   -1.767e+01  3.820e+02  -0.046 0.963107

    Null deviance: 1501.36  on 1082  degrees of freedom
Residual deviance:  896.01  on 1079  degrees of freedom
```

$$P(event\ is\ a\ hospital\ discharge) = \alpha + TimeDiff + Age + Frequency$$

α = A constant, in this case 18.31
Age = The age of the patient
TimeDiff = The amount of time (in seconds) it takes for the event to occur after the patient is admitted
Frequency = The number of times the event occurs for that particular patient

The **null deviance** shows the fit of the data to a null model, or a model containing only an intercept. The **residual deviance** shows the data's fit to a model with an intercept and variables. The smaller of these two shows if the model is useful or not. Because the residual deviance is smaller, we know that our model fits the data, and that the **probability of a hospital discharge is better predicted by our model than a null model**.

- In ROC curve analysis, an **ROC curve** is often used to determine the cutoff point for classification- i.e. the probability cutoff that **determines which probabilities obtained by a logistic regression produce one class (1), and which produce the other class (0)**.
- In the context of this problem, we are looking for the number that will be the cutoff point for determining which events are **hospital discharges** ("Release A", "Release B", "Release C", "Release D", or "Release E"), and which are not hospital discharges (all other events).
- In this case, the optimal cutoff will be the probability cutoff that gives the **smallest classification error.** The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
- The probability cutoff that gives the lowest misclassification error is **0.6973357**, so all regression outputs greater than this number will be classified as a **patient discharge**, and the outputs smaller than this number will be classified as **not a patient discharge**



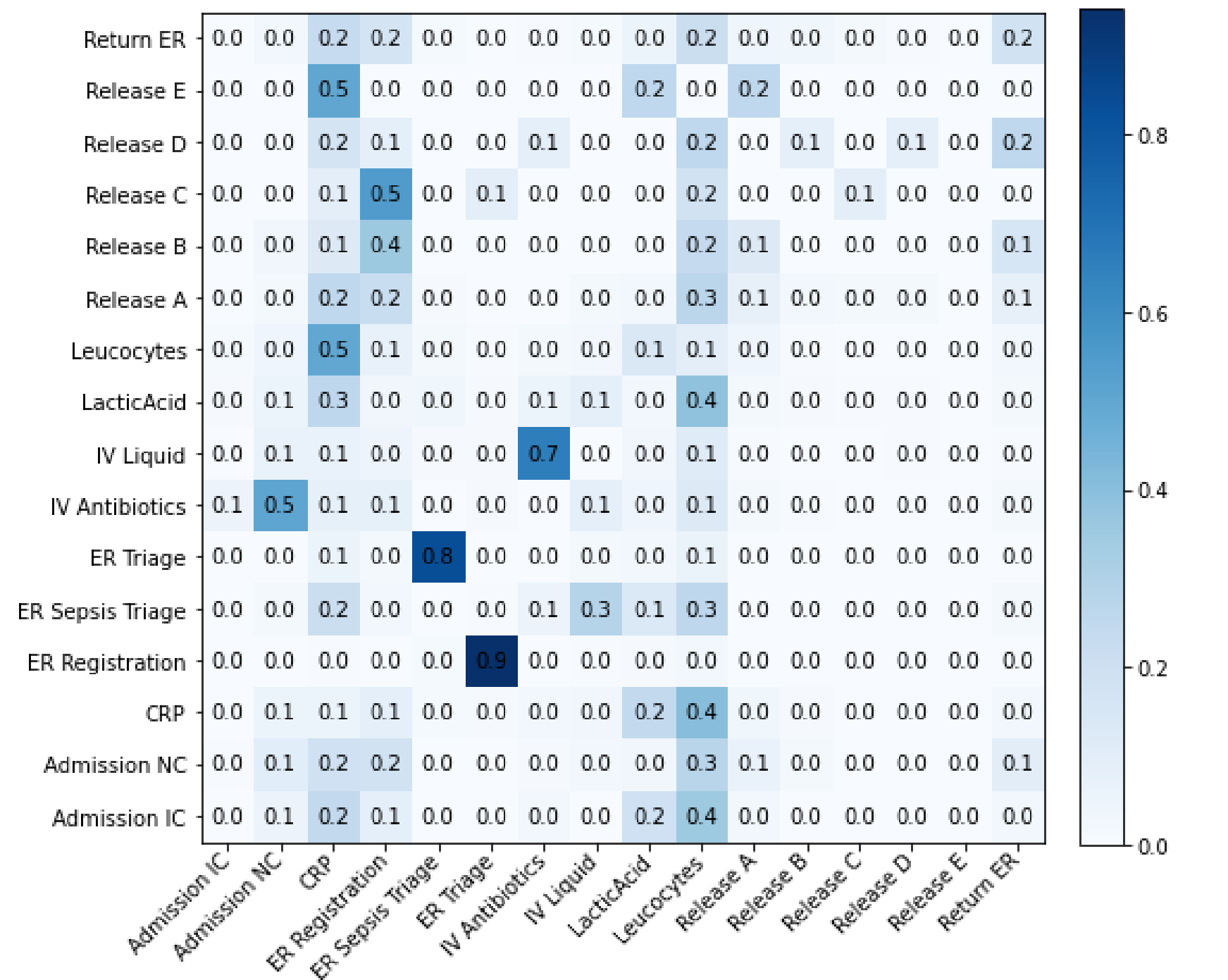
Markov Chains:

- **Sequences** of events where the probabilities of the future depends on the present.
  - **Cluster 1: Patient was not readmitted to hospital**
  - **Cluster 2: Patient was readmitted to hospital**
- By looking at each patient's sequence of events, we know that a patient was not readmitted to the hospital if their last event is "Release A". If the last event of their sequence is not "Release A", then they had to be **readmitted for additional treatments**.
- Compared and complemented the Precedence Diagram generated in Process Mining with the Markov Chains Transition Probability Matrices
- Used **probabilistic methods** to cluster patient event sequences into either the **"Readmitted to hospital"** or "**Not readmitted to hospital"** clusters
- Clustering model achieved **80.6% accuracy** on test event sequences

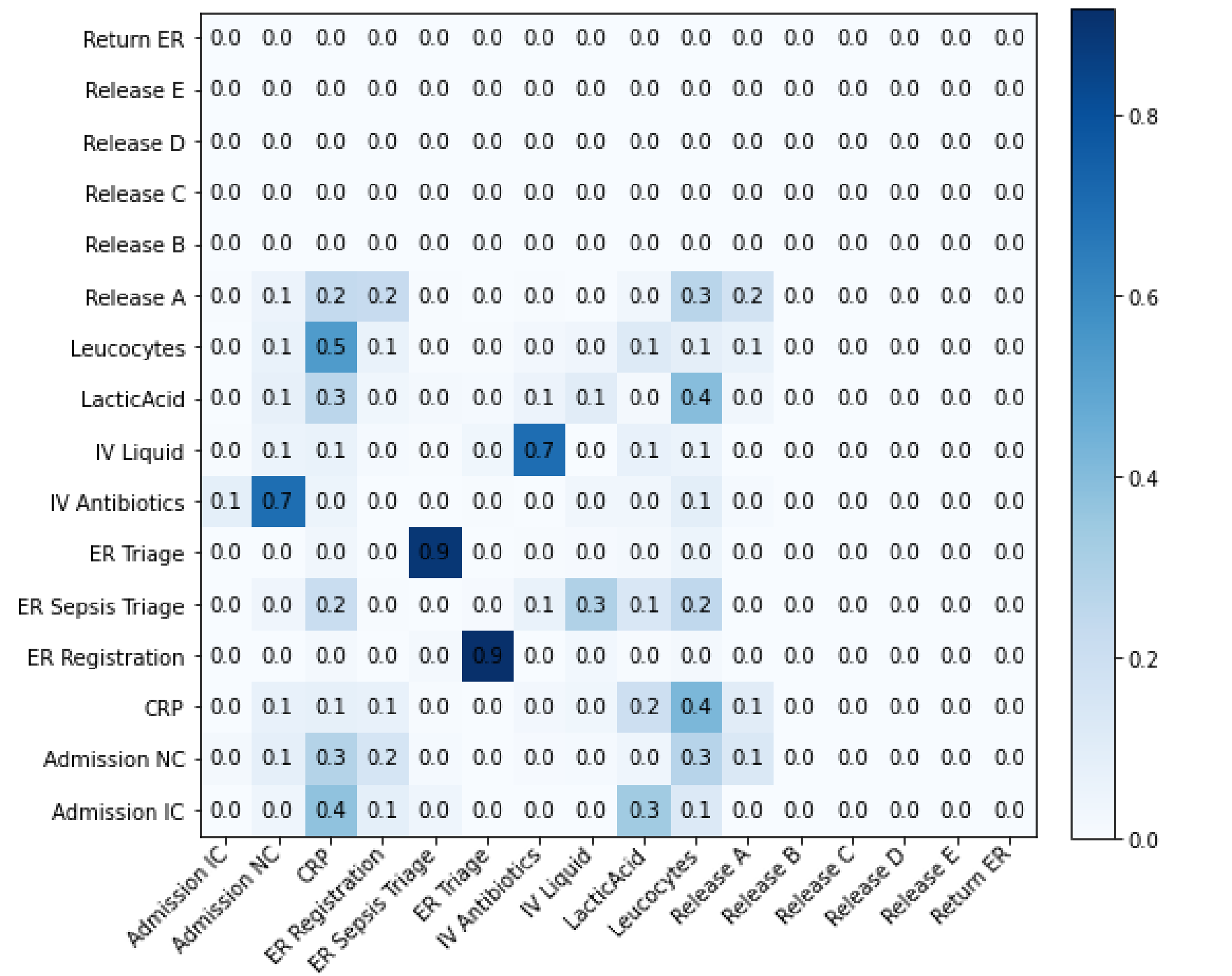
$$\mathrm{P(hospital\ readmission)} = \prod_{i=0}^{n} A[event_i, event_{i+1}]$$

$$\mathrm{P(no\ hospital\ readmission)} = \prod_{i=0}^{n} B[event_i, event_{i+1}]$$

Hospital Readmission Transition Probability Matrix (A)



No Hospital Readmission Transition Probability Matrix (B)



CONCLUSIONS

- The focus of this analysis was to better understand adult sepsis events and to develop machine learning models to predict patient discharge.
- The data set was analyzed and visualized to obtain an understanding of its organization and the distributions of its variables and components.
- Process model maps illustrating the different types of hospital events were generated to visualize the event log data and gain an understanding of the order and timely way requests are processed. From log files, we discovered the sepsis treatment process and the relationship between patient age and frequency of sepsis cases.
- A logistic regression model was then trained and tested to accurately predict which hospital events are discharges, and which are not.
- Lastly, Markov chain models were developed to perform sequence clustering to determine whether a patient will need to be readmitted to the hospital for additional treatments.
- Further work can be done to verify and fine-tune the models to obtain a higher classification accuracy. Additionally, other artificial intelligence methods can be implemented to provide other predictions and insights regarding event log data and processes.

ACKNOWLEDGEMENTS

- This project was supported in part by an appointment to the Science Education and Workforce Development Programs at Oak Ridge National Laboratory, administered by ORISE through the U.S. Department of Energy Oak Ridge Institute for Science and Education.
- We appreciate the helpful discussions about machine learning approaches with Dr. Hong-Jun Yoon.