Sepsis Cases

Seyedehsara Mirsepassi

30407A

Business Information Systems

Prof. Paolo Ceravolo

a.a. 2023 - 2024

Description of the case study

This real-life event log contains events of sepsis cases from a hospital. Sepsis is a life threatening condition typically caused by an infection. One case represents the pathway through the hospital. The events were recorded by the ERP (Enterprise Resource Planning) system of the hospital. There are about 1000 cases with in total 15,000 events that were recorded for 16 different activities. Moreover, 39 data attributes are recorded, e.g., the group responsible for the activity, the results of tests and information from checklists. Events and attribute values have been anonymized. The time stamps of events have been randomized, but the time between events within a trace has not been altered.

Organizational goals

Strategic Goals: Strategic goals are long-term objectives that focus on the overall improvement of healthcare outcomes for sepsis patients:

- Analyze Existing Data to get Valuable Insights into Patient Trajectories.
- Assess the Quality of the services provided by different departments of the hospital.
- Analyze Segments to Verify Correlations and evaluate Compliance with Sepsis Treatment Guidelines.
- Evaluate and Enhance the Effectiveness of Sepsis Treatment within the organization.

Tactical Goals: Tactical goals are medium-term actions designed to implement strategic goals effectively. They focus on specific projects or initiatives. For example, implement a data mining model to predict sepsis onset within 24 hours of patient admission.

Operational Goals: Operational goals are short-term, specific actions that support tactical goals and ensure daily activities are aligned with larger objectives: Optimize data collection processes to ensure high-quality input for the sepsis prediction model.

Knowledge Uplift Trail

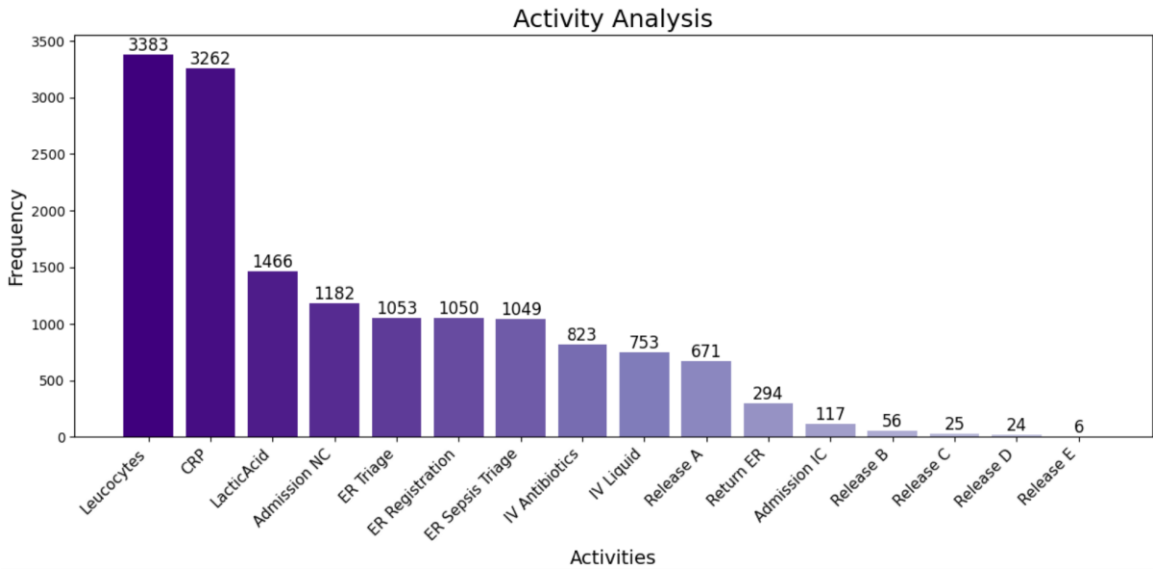
The Knowledge Uplift Trail (KUT) is a way to transform raw data into valuable information. It consists of several steps that help to better understand the basic features of the data, making it possible to gain insights from it. In this project below steps are used:

1. **Dataset Exploration:** Analyse the eventlog to extract knowledge related to our domain.
2. **Data Cleaning:** Prepares the event log by removing missing values and converting categorical variables to numerical formats.
3. **Data Filtering:** Establishes conditions to filter the data, retaining only cases that meet specific criteria according to hospital protocols.
4. **Variant Analysis:** Examines variations within the data to identify most common behaviour in the system.
5. **Process Mining:** Utilizes process mining techniques to discover and map the actual process flow.
6. **Conformance Checking:** Identifies any deviations from the standard sequence of events to ensure adherence to clinical guidelines.
7. **Descriptive Analysis:** Employs statistical tools to understand the distribution and characteristics of the data.
8. **Performance Analysis:** Analyze performance, workload and bottlenecks in the process.

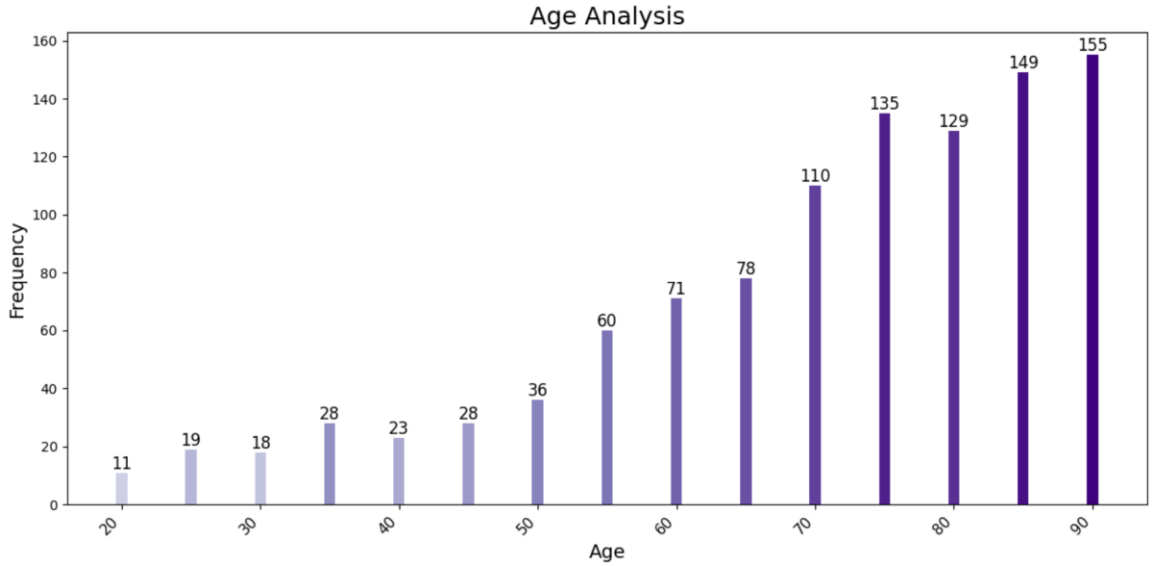
Dataset Exploration

This section gives an overview of the original dataset used in the study, focusing on some key attributes:

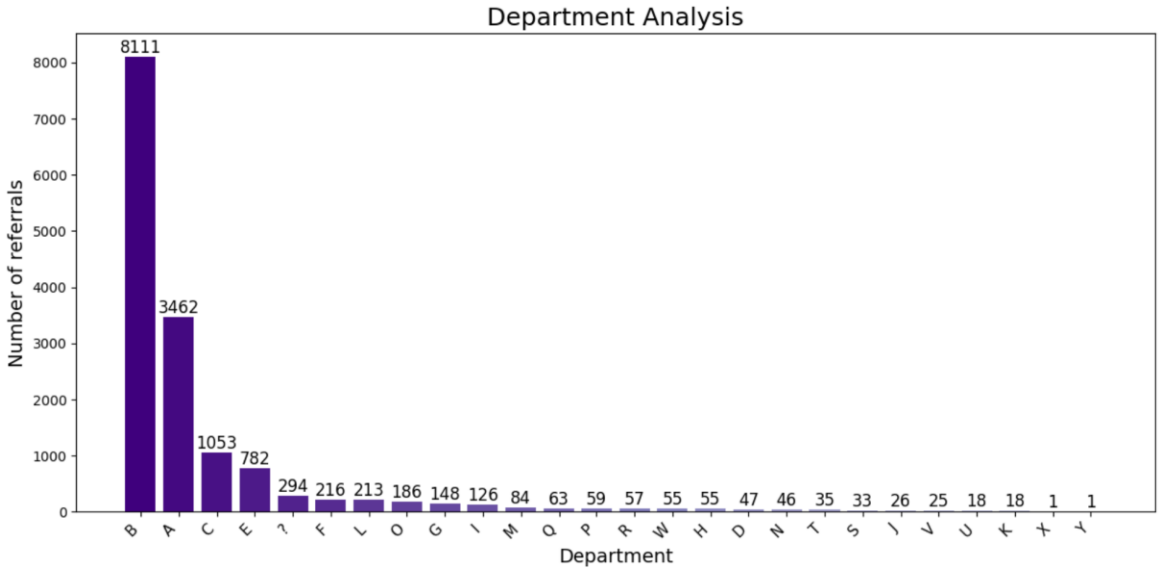
- **Event and Cases:** The dataset includes 1,050 cases and 15,214 events.
- **Activities:** There are 16 recorded activities. The most common are "Leucocytes", "CRP", and "Lactic Acid" which likely represent recurring blood tests. Less frequent activities like "Return ER" might indicate patient discharge:



- **Patient Age:** The average age of the patients is about 70, with more cases in older patients:

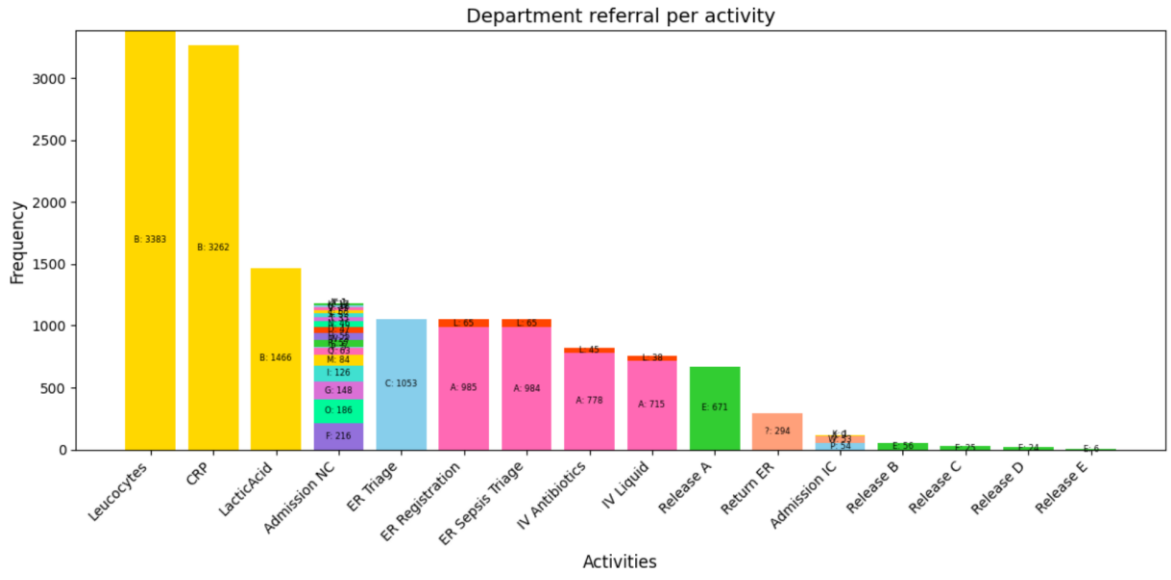


- **Departments:** The "org:group" column may represent different hospital departments. Most events are handled by departments "B", "A" and "C":

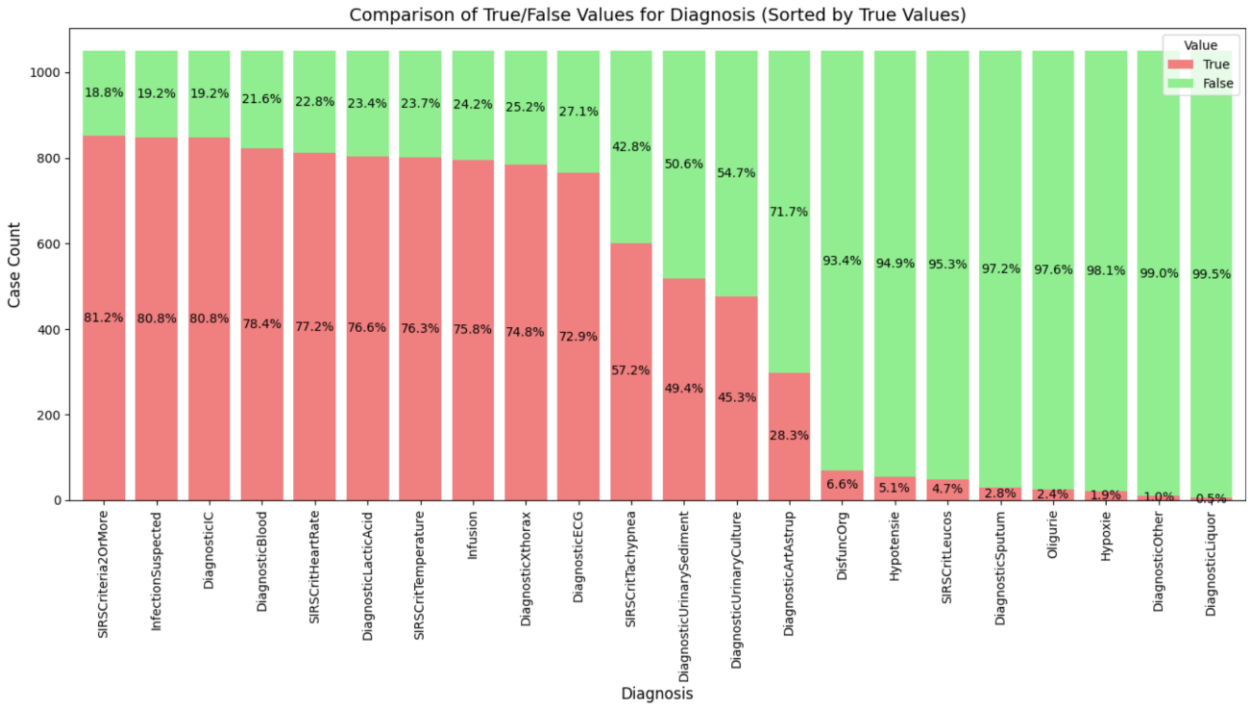


By analyzing the information related to executed events and involved departments, we can determine that

- Group "B" performs crucial tests like "Leucocytes", "CRP" and "Lactic Acid".
- Group "E" handles "Release" activities.
- Group "?" is involved in "Return ER."
- Group "C" manages "ER Triage."
- Group "A" is responsible for "ER Sepsis Triage," "ER Registration," "IV Liquid," and "IV Antibiotics."



The remaining attributes are mostly associated with diagnostic testing. As shown below, it appears that cases suspected of sepsis mostly have positive results in two or more criteria:



Diagnosis attribute distribution

Additionally, there are three more attributes related to 'Leucocytes', 'CRP' and 'LacticAcid' test results and will be discussed in more detail later.

Data Cleaning

As mentioned before, the dataset includes 1,050 potential patients and 15,214 events. Due to the critical role of 'Leucocytes', 'CRP', and 'LacticAcid' in sepsis analysis, we only consider cases that have values for these tests. Cases with missing or zero counts for these tests will be removed, resulting in the removal of 214 cases from the event log:

- ✓ Number of cases with missing or zero value in Leucocytes_events_count: 39
- ✓ Number of cases with missing or zero value in CRP_events_count: 103
- ✓ Number of cases with missing or zero value in LacticAcid_events_count: 193
- ✓ Total number of cases with missing or zero value in any of the count columns: 214

After filtering above cases, eventlog will contain 13522 events and 836 cases.

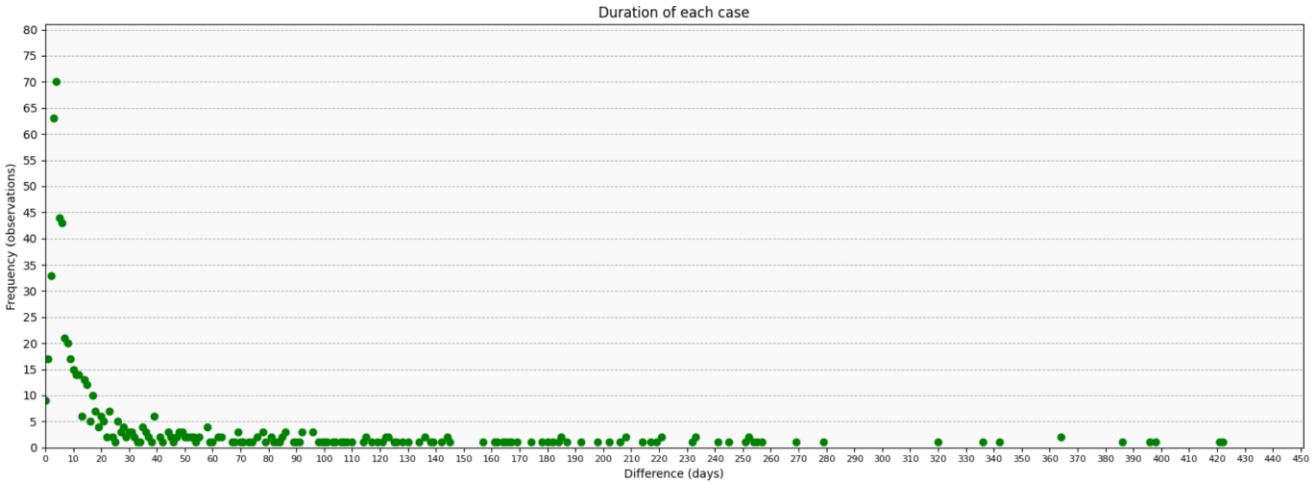
Data Filtering

In the next phase, the initial and final activities will be examined to identify complete cases. Potential starting activities could include 'ER Registration', 'ER Sepsis Triage' or 'ER Triage'. Possible concluding activities might be 'Release A', 'Release B', 'Release C', 'Release D', 'Release E' or 'Return ER'. These constraints will maintain the accuracy of the sepsis case analysis:

- ✓ **Number of events after filtering:** 11566
- ✓ **Number of cases after filtering:** 655
- ✓ **Start activities after filtering:** {'ER Registration': 646, 'ER Sepsis Triage': 5, 'ER Triage': 4}
- ✓ **End activities after filtering:** {'Release A': 327, 'Return ER': 257, 'Release B': 44, 'Release E': 3, 'Release C': 14, 'Release D': 10}

In the next step, the case durations also the cumulative times for each event per case will be calculated for further analysis:

- ✓ **Mean of the differences:** 39 days 08:37:54.983206107
- ✓ **Median of the differences:** 9 days 17:44:14
- ✓ **Mode Duration:** 11666000000000 nanoseconds
- ✓ **Standard Deviation:** 68 days 22:38:32.899655564



As shown in the graph above, most cases are completed in less than a month. However, some cases take significantly longer, requiring further analysis to understand these extended durations and find the probable bottlenecks. Therefore no filtering will be applied based on case durations.

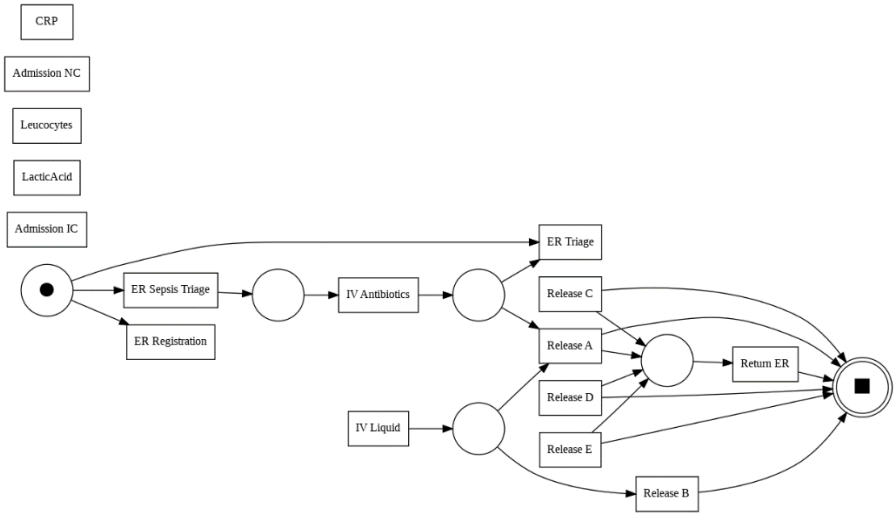
Variant Analysis

By analysing the filtered 655 cases, it was noticed that many sequences appeared only once. The most common sequences only appeared four times, which is a small number compared to the total cases. This highlights how complex the process is and how hard it is to find a small group of common sequences that cover most cases. Below are top variants:

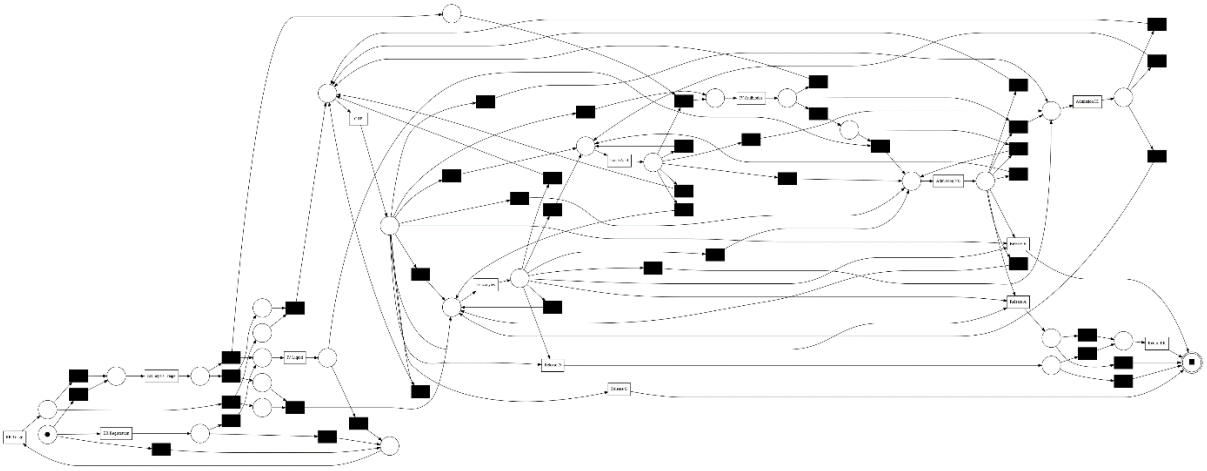
- **Number of cases of trace ('ER Registration', 'ER Triage', 'ER Sepsis Triage', 'LacticAcid', 'Leucocytes', 'CRP', 'IV Liquid', 'IV Antibiotics', 'Admission NC', 'Release A'): 4**
- **Number of cases of trace ('ER Registration', 'ER Triage', 'ER Sepsis Triage', 'Leucocytes', 'LacticAcid', 'CRP', 'IV Liquid', 'IV Antibiotics', 'Admission NC', 'CRP', 'Leucocytes', 'Release A'): 4**
- **Number of cases of trace ('ER Registration', 'ER Triage', 'ER Sepsis Triage', 'Leucocytes', 'CRP', 'LacticAcid', 'IV Liquid', 'IV Antibiotics', 'Admission NC', 'Leucocytes', 'CRP', 'Release A'): 4**
- **Number of cases of trace ('ER Registration', 'ER Triage', 'ER Sepsis Triage', 'IV Liquid', 'Leucocytes', 'CRP', 'LacticAcid', 'IV Antibiotics', 'Admission NC', 'Leucocytes', 'CRP', 'Release A', 'Return ER'): 3**
- **Number of cases of trace ('ER Registration', 'ER Triage', 'ER Sepsis Triage', 'Leucocytes', 'CRP', 'LacticAcid', 'IV Liquid', 'IV Antibiotics', 'Admission NC', 'CRP', 'Release A', 'Return ER'): 3**
- **Number of cases of trace ('ER Registration', 'ER Triage', 'ER Sepsis Triage', 'CRP', 'LacticAcid', 'Leucocytes', 'IV Liquid', 'IV Antibiotics', 'Admission NC', 'Release A'): 3**
- **Number of cases of trace ('ER Registration', 'ER Triage', 'ER Sepsis Triage', 'IV Liquid', 'IV Antibiotics', 'Leucocytes', 'LacticAcid', 'CRP', 'Admission NC', 'Leucocytes', 'CRP', 'Release A', 'Return ER'): 3**

Process Mining

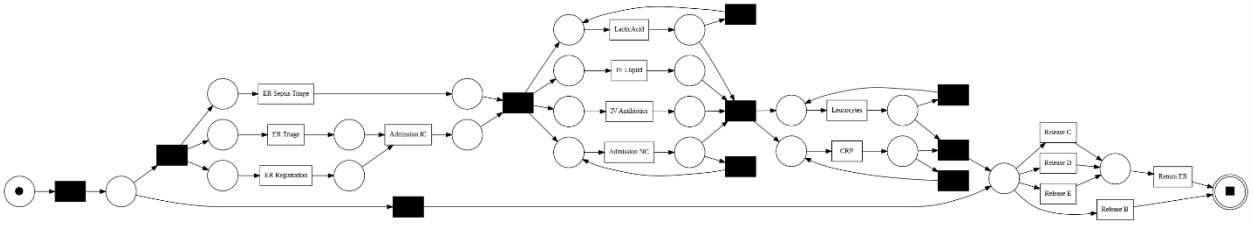
While having such variability in variants, in next step I try to use process discovery algorithms such as Heuristics Miner, Alpha Miner and Inductive Miner algorithms to build Petri nets from the event log. Then, I will evaluate two sample Petri nets based on their fitness, precision, generalization and simplicity.



Petri net - Alpha Miner



Petri net - Heuristics Miner (dependency_threshold=0.9)



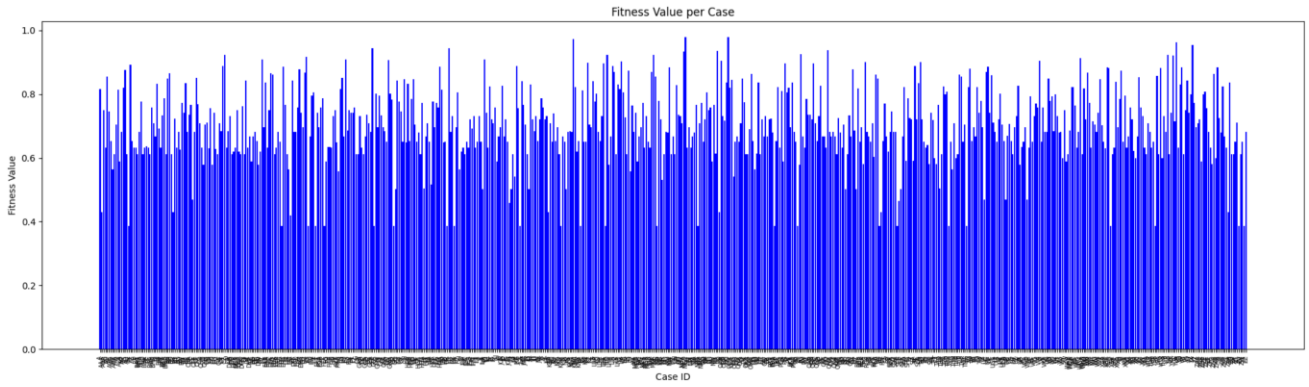
Petri net - Inductive Miner (noise_threshold=0.9)

	Fitness avoid non-fitting behavior	Precision avoid under-fitting behaviour	Generalization avoid overfitting behaviour	Simplicity
Alpha Miner	0.7428250281004121	0.19249965969417848	0.8996666814581679	1.0
Inductive Miner	0.6269963017401141	0.4357696299693451	0.8772059559036053	0.6533333333333333

Comparing Quality Metrics

Conformance Checking

To assess the alignment of each trace with the process model (Inductive Miner process model is considered), token-based replay conformance checking is used. As expected due to high variability in variants, cases where not fitted well to the process model:



Fitness Value Per Case

To improve the log slightly, I decided to remove the cases with fitness values less than 0.5. As a result, 38 cases were removed from the log:

['AAA', 'AV', 'BR', 'CE', 'EH', 'EJA', 'ET', 'F', 'FEA', 'GFA', 'GP', 'HR', 'HX', 'JGA', 'JL', 'KB', 'KHA', 'KP', 'LFA', 'MO', 'NKA', 'O', 'PEA', 'PS', 'RS', 'RT', 'SE', 'SEA', 'TDA', 'TMA', 'UA', 'UP', 'VD', 'XD', 'YEA', 'ZLA', 'ZU', 'ZX']

Again, the quality Matrix were calculated but no significant difference is observed:

	Fitness avoid non-fitting behavior	Precision avoid under-fitting behaviour	Generalization avoid overfitting behaviour	Simplicity
Alpha Miner	0.7449524445186051	0.19250854938427586	0.8488811844981619	1.0
Inductive Miner	0.6999831430919659	0.3057437756066814	0.85565393852869	0.641025641025641

Comparing Quality Metrics (After removing non-conformant cases)

Descriptive Analysis

In the first step, high-risk cases are analysed based on the following criteria:

- A WBC counts of less than 4,000/µl or more than 12,000/µl.
- LacticAcid levels > 2.0 µg/L indicate a high probability of systemic bacterial infection and a risk of progression to sepsis or septic shock. LacticAcid levels < 0.5 µg/L indicate a low probability of systemic bacterial infection and a low risk of progression to sepsis or septic shock.
- A plasma CRP of 50 mg/l or more was highly suggestive of sepsis. The risk of sepsis-related mortality appears to be increased when the 3rd day CRP value is greater than 100 mg/L.

The result of analysis on 607 cases is as follows:

- **CRP more than 50 MG/L:** 556 Cases
- **Leucocytes less than 4,000/µl or more than 12,000/µl:** 442 Cases
- **LacticAcid more than 2.0 MG/L:** 249 Cases
- **Total umber of cases which have at least one symptom:** 592 Cases

In the next step, Lift value can help us understand how each segment's traits are linked to the likelihood of a "Return ER." To do this, we need to count the number of "return_er" events in each segment.

$$lift(x \rightarrow y) = \frac{supp(x \rightarrow y)}{supp(x) \times supp(y)}, range : [0,\infty]$$

Start activities_suspicious_cases: {'ER Registration': 583, 'ER Sepsis Triage': 5, 'ER Triage': 4}

End activities_suspicious_cases: {'Release A': 289, 'Return ER': 232, 'Release B': 44, 'Release E': 3, 'Release C': 14, 'Release D': 10}

Start activities_CRP: {'ER Registration': 548, 'ER Sepsis Triage': 5, 'ER Triage': 3}
End activities_CRP: {'Release A': 274, 'Return ER': 214, 'Release B': 42, 'Release E': 3, 'Release C': 13, 'Release D': 10}

Start activities_Leucocyte: {'ER Registration': 436, 'ER Triage': 3, 'ER Sepsis Triage': 3}
End activities_Leucocyte: {'Release A': 207, 'Return ER': 179, 'Release B': 32, 'Release E': 3, 'Release C': 12, 'Release D': 9}

Start activities_Lactic: {'ER Registration': 244, 'ER Sepsis Triage': 3, 'ER Triage': 2}
End activities_Lactic: {'Release A': 114, 'Return ER': 102, 'Release B': 25, 'Release E': 1, 'Release C': 2, 'Release D': 5}

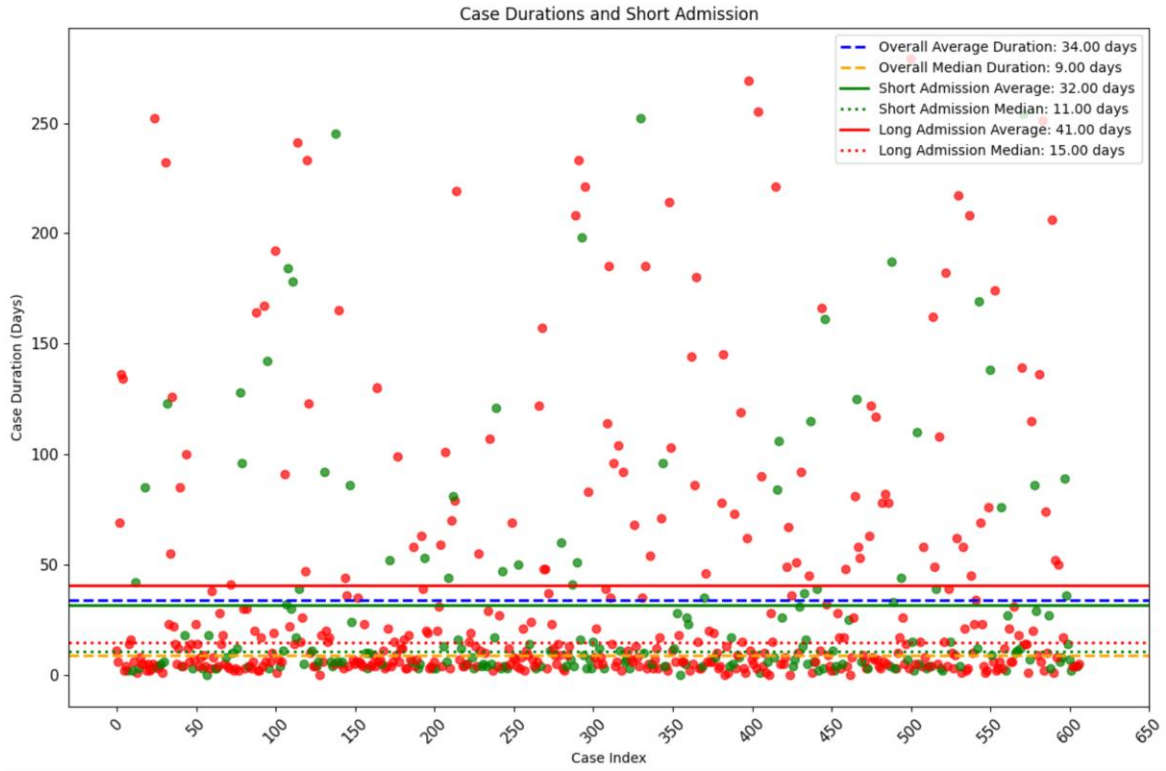
	CRP Positive	Leucocyte Positive	LacticAcid Positive
Lift Value	1.045730123913769	1.3840886931485632	2.4851745995198073

Lift Value Comparision

It can be seen that the cases with LacticAcid abnormality are more probable to return.

For the next evaluation, liquid and Antibiotic administration are analyzed. Following medical guidelines the quick administration of Liquids and antibiotics is crucial in Sepsis cases. Therefore, the we investigated the durations between the case start time to the administration of IV antibiotics and fluids, focusing on whether it is less than one hour or more:

- ✓ Given 607 number of cases with high risk of Sepsis, there are 200 cases where their first administration of 'IV Antibiotics' is less than 1 hour.
- ✓ Given 607 number of cases with high risk of Sepsis, there are 271 cases where their first administration of 'IV Liquid' is less than 1 hour.
- ✓ There are 186 cases that meet all 2 medical guidelines for sepsis treatment.



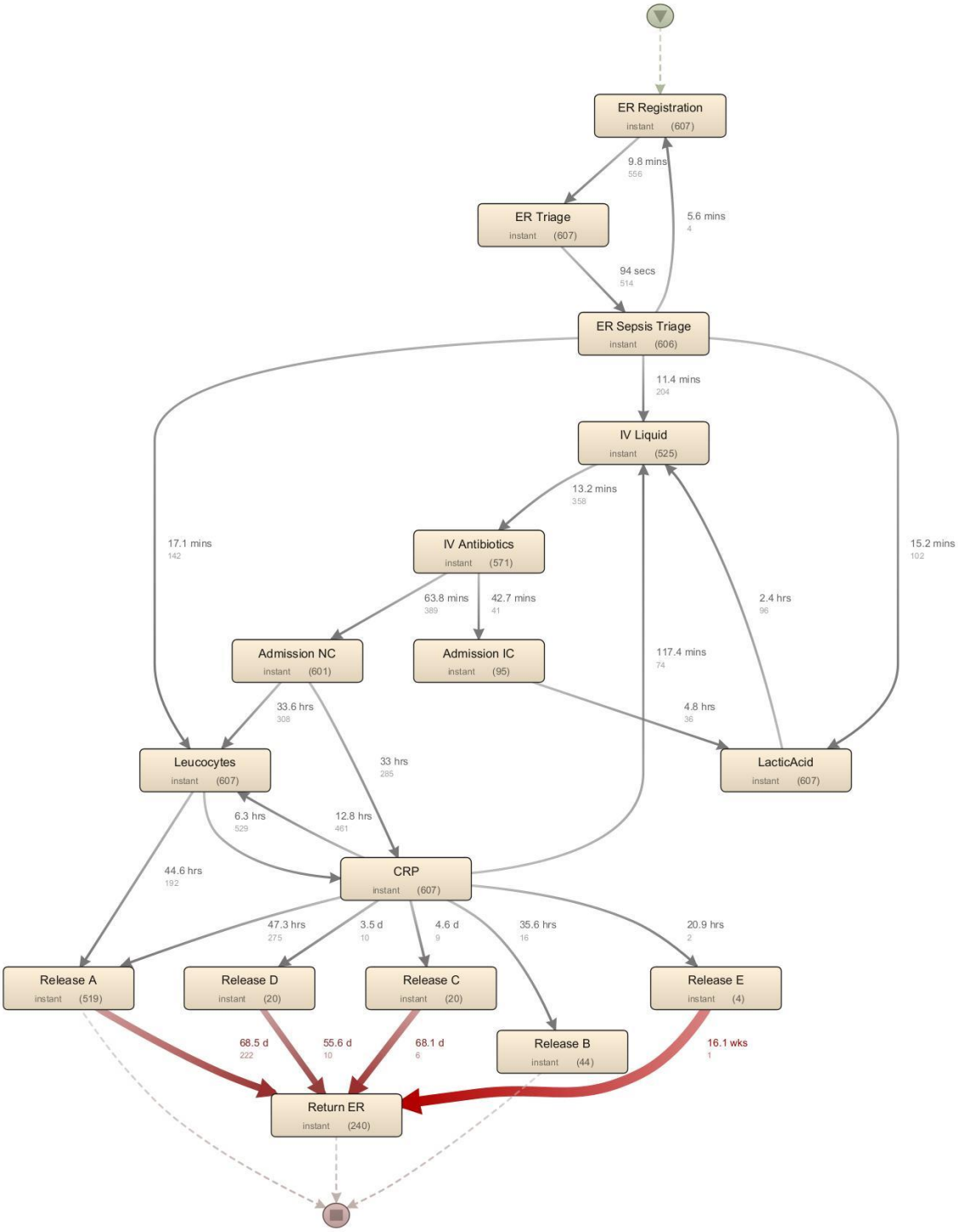
Case Duration Comparision according to Liquid/Antibiotic administration time

As shown in the graph and table, the average case duration for patients who received fluids and antibiotics within an hour of arriving at the hospital is much shorter than for those who did not. So we can consider it for further improvements:

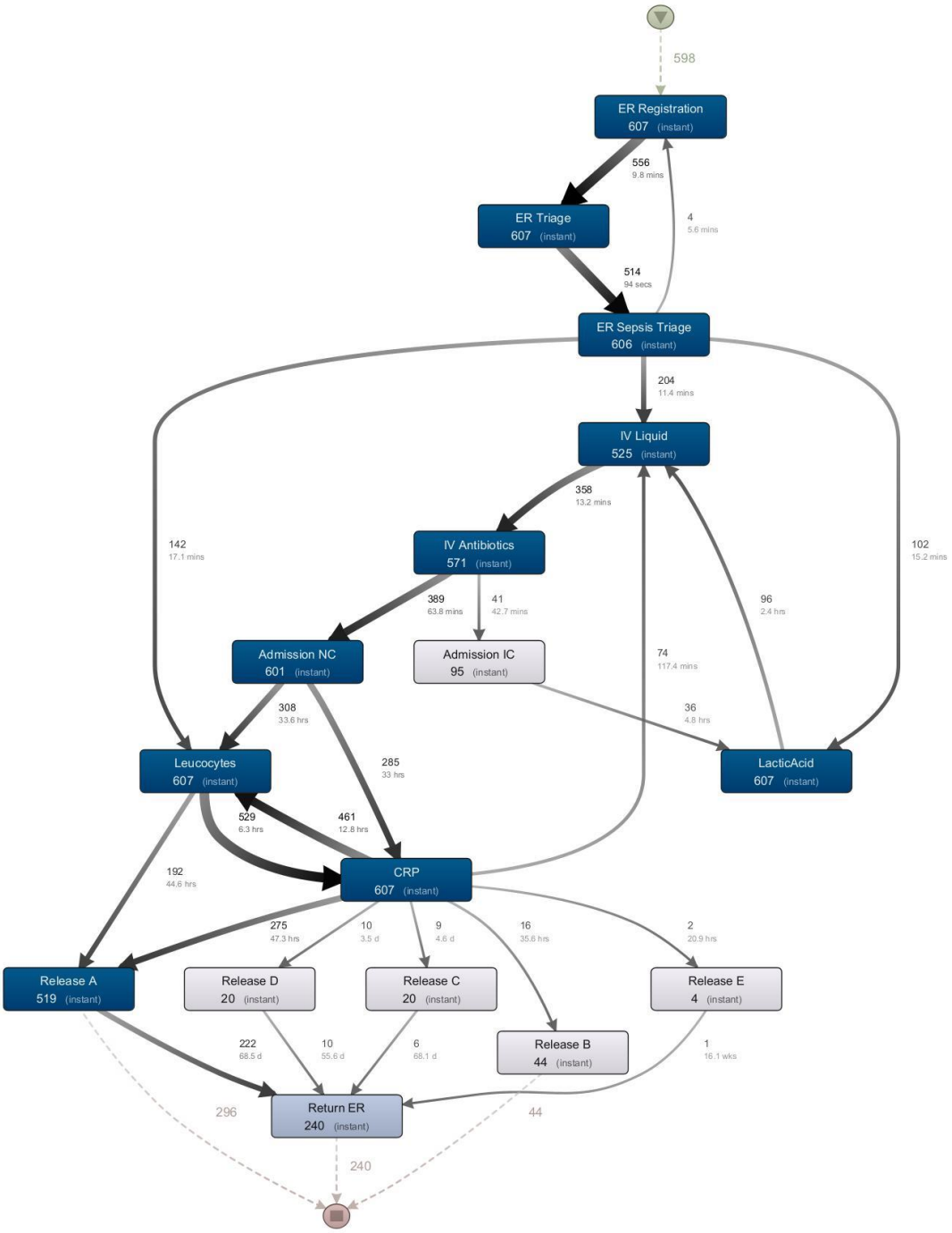
	Mean	Median	Mode	Standard Deviation
Administration done in less than 1 hour	32	11	29	49
Administration done in more than 1 hour	41	15	114	56

Performance Analysis

Below graphs exported from Disco software give us a clear view of the system's bottlenecks and workload. It's clear there are major bottlenecks between the release points and the returner that need further investigation to improve the process:



Identifying Bottleneks



System workload

Conclusion

During this project, we took steps toward the goals defined at the beginning:

- Analyze Existing Data to get Valuable Insights into Patient Trajectories.

By analyzing existing data, valuable information about patient trajectories were provided from start to discharge. This includes assessing risky cases, evaluating how treatments impact patient recovery, understanding system performance and identifying bottlenecks.

- Assess the Quality of the services provided by different departments of the hospital.

By identifying which departments are involved at each stage of patient trajectory and evaluating their workload, this assessment helps to highlight areas for potential improvement.

- Analyze Segments to Verify Correlations and evaluate Compliance with Sepsis Treatment Guidelines:

Analyzing cases with abnormalities In CRP, Leucocytes, and LacticAcid helps us to find out that LacticAcid levels have a significant impact on patient returns to emergency rooms. Also rapid admission of antibiotics and liquids significantly reduces case durations.

- Evaluate and Enhance the Effectiveness of Sepsis Treatment within the organization.

To enhance treatment effectiveness in future, below Items are suggested:

- According to the provided workload and department involvement in sepsis cases, consider additional resources or automation for quicker responses.
- Address identified bottlenecks to reduce treatment duration.
- Considering the higher Lift value related to the cases with LacticAcid abnormalities, any improvement in monitoring and treatment of this symptom can be benefitial.
- Implement protocols for rapid Antibiotic and Liquid admission to improve patient recovery in the future.

Reference

Source code is available at**:**

https://github.com/saramirs/Business-Information-System

https://colab.research.google.com/drive/192CpYA2oOD03mipOV4GvFUMcSP7AVURt