## Introduction

This report presents a comprehensive analysis of the handling of patients suspected of sepsis in one of the Dutch hospitals. Sepsis, a critical medical condition, is defined as the body's extreme response to an infection and is potentially life-threatening if not promptly diagnosed and treated. The significance of this study lies in its potential to improve patient care and outcomes in sepsis cases.

According to the research done on the subject [1], process mining techniques can provide vast and valuable insights into the trajectory of patients in a hospital setting. This innovative approach allows researchers to map out the complex journey of sepsis patients from admission to discharge. Moreover, it enables the verification of whether the daily clinical practice in the hospital aligns with established medical guidelines for sepsis management. This comparison between actual practice and guidelines is crucial for identifying areas of improvement in patient care protocols.

In the research that will be presented below, we performed several processes to gain insights into the process of treating sepsis patients. We started by understanding the database and cleaning it, from there we moved on to discovering the model and performing conformance checking against the database data, while choosing the optimal model and checking the processes that the patients go through in relation to it, and whether there is anything to improve in the existing processes in the organization. Finally, we tried to improve the model we received using additional techniques, for example: checking the time for a process that a certain patient goes through, mining decision points, and more.

## A literature survey on process mining techniques and their use in our project

In this section, we present methods used during our project at different stages, such as discovering the model, performing conformance checking and improving the model. This literature survey examines four key process mining techniques: Alpha Miner, Inductive Miner, Conformance Checking, and Decision Point Mining.

The Alpha Miner [2] constructs a Petri net representation of a process by identifying direct succession, causality, and parallel relationships between activities in event logs. The algorithm examines the order of events in the log and infers the underlying process structure. While providing a basic understanding of process structure, it has limitations in handling complex behaviors such as loops and noise in real-world logs. We used Alpha Miner to create a basic model, from which we could advance to better discovery methods.

The Inductive Miner [3] addresses some of the Alpha Miner's limitations. It uses a divide-and-conquer approach to recursively split the event log and discover process fragments, which are then combined into a process tree. This technique is effective in handling large and complex event logs, making it suitable for analyzing intricate processes. We used inductive miner to more accurately discover our model, tested its soundness, and used the resulting model as we'll explain in the report.

Conformance Checking [4] evaluates the alignment between process models and actual process executions. It compares the discovered process model with a predefined reference model or set of rules. This technique allows organizations to assess compliance with procedures, identify deviations, and measure the fitness of their process models. We used conformance checking for two purposes: to determine which model best fits our data, and mainly to examine an existing model's suitability against actual hospital behavior. This allowed us to map non-conformities and identify less suitable process parts, revealing opportunities for improvement.

Decision Point Mining [5] analyzes choices made at specific points in a process. This technique uncovers factors influencing decisions, providing insights into rules and patterns governing these choices. By examining case characteristics at decision points, it helps identify criteria used in process-related decisions. We tried to use this and other methods to try to improve our discovered model.

**Description of the database and preprocessing the database**

Our study examines sepsis cases in a Dutch emergency ward, utilizing a robust event log collected over 18 months from November 2013 to June 2015. The dataset encompasses 1,050 patient traces, comprising 15,214 events across 16 activity types and 39 data attributes per case. By integrating information from various hospital systems, including triage documents, blood tests, and financial records, the study enables a detailed analysis of patient journeys. Applying process mining techniques to this rich dataset allows researchers to assess guideline adherence, identify treatment protocol deviations, visualize care pathways, and analyze patterns of patient returns within 28 days. Ultimately, our research aims to enhance emergency care processes and improve patient outcomes by providing valuable insights into the complexities of sepsis treatment in an emergency setting .the possible steps in the process include:

1. Registration in the emergency room: This is the initial step where patients with symptoms of sepsis are registered upon arrival at the hospital's emergency ward.
2. Triage: A triage document is filled out for sepsis patients, which includes information such as the time of triage, symptoms present (SIRS criteria for sepsis), diagnostics ordered, and the time infusions of liquid and antibiotics were administered.
3. Medical activities: These activities include measurements of leukocytes, CRP (C-reactive protein), and lactic acid. These tests are likely performed to assess the severity of the sepsis condition and guide treatment decisions.
4. Admission to hospital wards: Patients may be admitted to different wards based on their condition. The document mentions admission to normal care wards and intensive care wards as possible trajectories.
5. Discharge: There are different variants of discharge from the hospital, which may depend on the patient's condition and response to treatment.

6. Returning patients: The document mentions investigating the trajectory of patients who return to the hospital within 28 days. This suggests that some patients may require readmission or further treatment after their initial discharge.

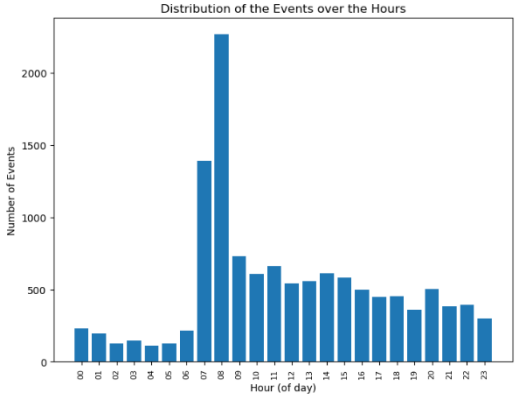
we highlighted the most common process variants observed in the patient handling process for sepsis cases:

| | variant | count | percentile |
|---|---|---|---|
| 21 | (ER Registration, ER Triage, ER Sepsis Triage) | 35 | 3.333333 |
| 24 | (ER Registration, ER Triage, ER Sepsis Triage, Leucocytes, CRP) | 24 | 2.285714 |
| 95 | (ER Registration, ER Triage, ER Sepsis Triage, CRP, Leucocytes) | 22 | 2.095238 |
| 3 | (ER Registration, ER Triage, ER Sepsis Triage, CRP, LacticAcid, Leucocytes, IV Liquid, IV Antibiotics) | 13 | 1.238095 |
| 36 | (ER Registration, ER Triage, ER Sepsis Triage, Leucocytes, CRP, LacticAcid) | 11 | 1.047619 |

As we can see, variants such as ER Registration, ER Triage, and ER Sepsis Triage were prevalent in the event log data, indicating the standard sequence of activities when managing sepsis patients. Interestingly, none of these common variants concluded with a discharge, suggesting a deviation from the expected endpoint in the patient trajectory. This discrepancy raised concerns about the completion of the patient handling process and the need for further investigation to address this issue.

As a result, we decided to filter our event log data using specific start and end activities. The valid start activity was ER Registration, marking the beginning of a patient's trajectory in the hospital. The valid end activities, including Release A, Release B, Release C, Release D, Release E, and Return ER, signified different points at which a patient's journey could conclude. By filtering the event log based on these start and end activities, a subset of 734 traces was isolated for further examination and process modeling. This filtering step was crucial in focusing the analysis on complete patient cases and understanding the flow of activities from registration to discharge in handling sepsis patients within the hospital environment (The most common traces after the filtering can be seen in th Appendices section).

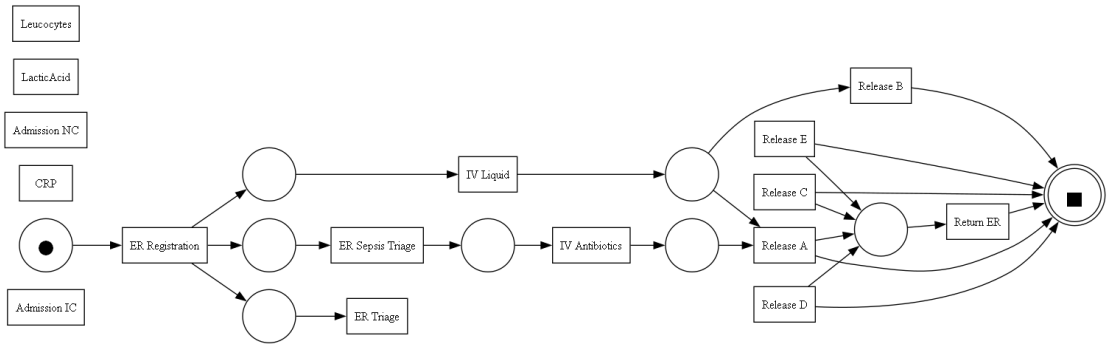
In addition, we also analyzed the distribution of events that occur in varying periods of time (by month, by day, and by hour), and also the distribution of events by the total duration recorded in the database (Graphs for this section are in the Appendices section). We noticed that most events happen in the hospital in the morning. It can be concluded from this that people naturally come to the hospital more in the morning after they get up, even if something happens to them the night before.



## Process Discovery

During the model discovery process, we used various techniques and different discovery algorithms to discover our model. Our goal was to create a model that is easy to understand, work quickly, and give accurate results. To do this, we tried different ways to build the model. We experimented with using more or fewer steps in the process.

We started with a basic model discovery that was revealed by the alpha algorithm from all of the traces in the filtered log:

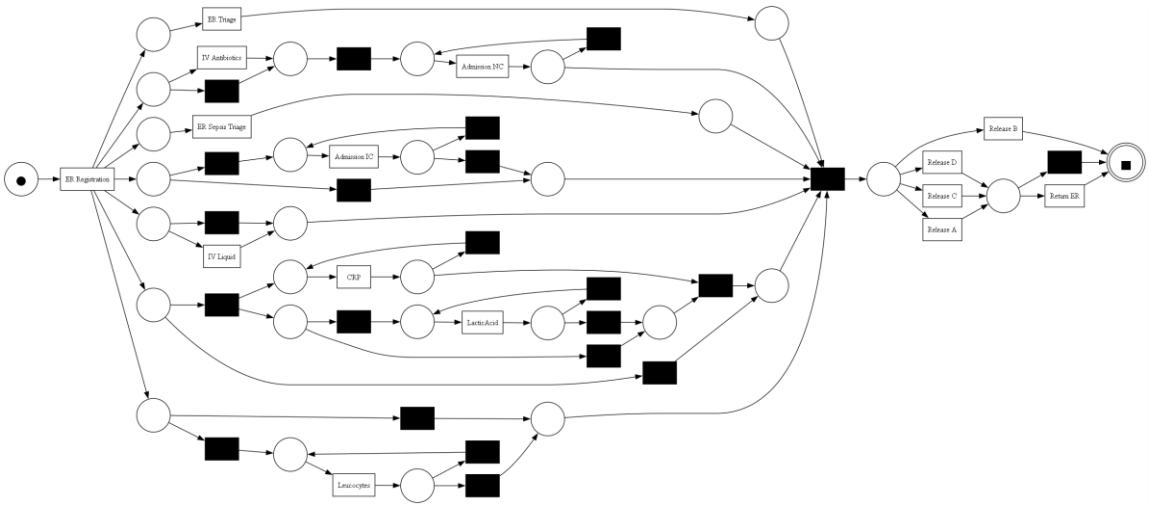



As we can see (and in accordance with what we explained in the literature review, as well as what was taught in the course), the algorithm produced a model that is not sound: Not all the parts are connected to each other as we would like to get. Therefore, we chose to focus on other algorithms for discovering our model.

As we continued the process, we chose to focus on different ways to discover our model. First, we used the inductive algorithm and the heuristic algorithm to discover our model instead of the alpha algorithm. During the process, we noticed that the heuristic algorithm also doesn't necessarily produce sound models, so we decided to proceed with the inductive algorithm, since as we saw in the lecture, inductive algorithm is the algorithm that gives the results that are sound and hence we chose our model from the results of the inductive algorithm (although as can be seen in the code, both methods we'll detail in the model discovery were implemented with both the inductive algorithm and the heuristic algorithm, where possible). The results we'll present refer to the execution of the inductive algorithm.

In the first stage, we randomly selected a number of traces each time and discovered the model from them (and as we mentioned above, we did this with both the inductive algorithm and (where possible, of course) with the heuristic algorithm). We performed conformance checking (Throughout the project, we performed conformance checking using the alignments method, because in the course we learned that this is the most correct method for performing conformance checking. Also, of course, the conformance checking was performed against the traces from the log from which we did not discover the model from, and as you can see in the code) for each result so that we could choose the best model. In each iteration, we increased the number of traces from which we discovered the model and

examined the results. For example, for 100 random traces using the inductive algorithm, the following model was discovered:




We noticed in the conformance checking results we received that the models obtained in this way present unsatisfactory results. As can be seen below for a varying number of traces from which we discovered the model, we received results that are very overfit. Additionally, in the discovered models, the simplicity measure is very low, and the precision measure is also very low, which indicates that the models can "invent" traces that don't exist in high quantities.

*50 random traces:*

| | Metric | Value |
|---|---|---|
| 0 | Fitness | 0.997606 |
| 1 | Precision | 0.504428 |
| 2 | Generalization | 0.924578 |
| 3 | Simplicity | 0.600000 |

*100 random traces:*

| | Metric | Value |
|---|---|---|
| 0 | Fitness | 0.997921 |
| 1 | Precision | 0.490565 |
| 2 | Generalization | 0.893889 |
| 3 | Simplicity | 0.614679 |

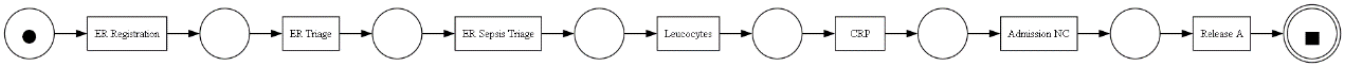
*200 random traces:*

| | Metric | Value |
|---|---|---|
| 0 | Fitness | 0.997959 |
| 1 | Precision | 0.420599 |
| 2 | Generalization | 0.893922 |
| 3 | Simplicity | 0.600000 |

We decided to take a different approach to get better results and find the best model for our data. Now we tried to build the model each time from a varying number of the most common traces (where each time we increased the number of most common traces from which we discovered the model, meaning for example, at first, we built the model from the most common trace, then from the three most common traces, and so on). It is important to note that this approach was taken after consultation with the course lecturer, and as you can see in the appendices, he approved the approach we took. In this way, we discovered that the best model, considering the conformance checking results (as explained in detail in the Conformance Checking section below. Also, in here too, the conformance checking was performed against the traces from the log from which we did not discover the model, and as you can see in the code), was obtained by applying the inductive algorithm the three most frequent variants in the model. Our model is sound (we checked it in the code), and The train-test ratio in that case was: train: 1.77% , test: 98.23%. Here too we consulted with the course lecturer about the train-test ratio we received, and he said that what is important is the results of the model and not this ratio, as you can see in the appendices.

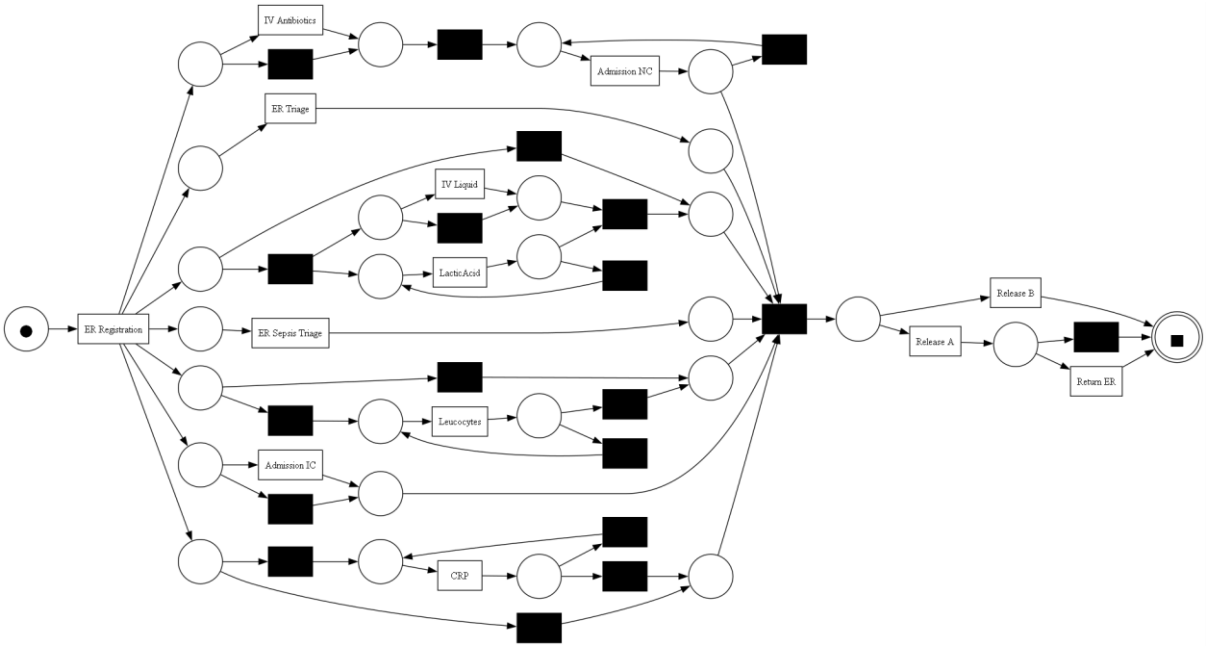
By limiting the model to these three variants, we maintained a high level of precision and simplicity while ensuring that the model accurately represents the key steps in the process flow of handling sepsis patients in the Emergency Room.

Using only the most common variant gave us this petri net-



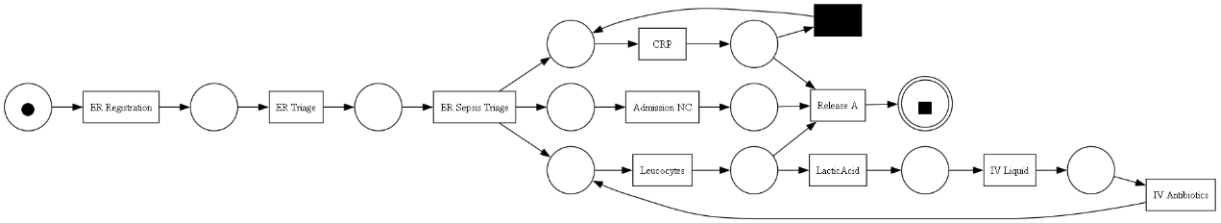

As we expected- one variant gives very simple petri net, but it doesn't fit the whole data.

However, when we used 50 variants, we got this-




Which is very complicated petri net, with much higher fitness (99.02%).

The 3 most common variants petri net-




we think that this model is a suitable model (As we will explain in detail in the Conformance Checking section). this model is simpler and faster, but it still gives us a good picture of how to care for sepsis patients in the emergency room.

**Conformance Checking**

Conformance checking serves multiple purposes, as discussed in our course. One primary goal is model discovery, which involves engineering a conformance checking algorithm to uncover a suitable process model. While there may not be a single "best" model, this approach allows for the identification of a model that accurately represents the observed behavior. The second and more

critical objective of conformance checking is to evaluate the conformance of an existing model. This evaluation is performed against a larger log of traces that were not used to generate the model. As we mentioned above, When we did conformance checking, it was performed against the traces from the log from which we did not discover the model. We evaluated four quality metrics for each model: Fitness, Precision, Generalization, Simplicity and we compared the results across different trace counts:

*1 most common trace:*

| | Metric | Value |
|---|---|---|
| 0 | Average Fitness | 0.586572 |
| 1 | Precision | 1.000000 |
| 2 | Generalization | 0.968781 |
| 3 | Simplicity | 1.000000 |

*3 most common trace:*

| | Metric | Value |
|---|---|---|
| 0 | Average Fitness | 0.766402 |
| 1 | Precision | 0.931965 |
| 2 | Generalization | 0.968088 |
| 3 | Simplicity | 0.793103 |

*5 most common trace:*

| | Metric | Value |
|---|---|---|
| 0 | Average Fitness | 0.882228 |
| 1 | Precision | 0.788969 |
| 2 | Generalization | 0.955331 |
| 3 | Simplicity | 0.674419 |

Among the three models, the model with three variants seems like the most effective choice. It strikes a balance between achieving high performance across all metrics and maintaining a manageable level of complexity and precision.

**Fitness**, which measures how well the model aligns with real-world data, reaches a 76.64% for our model. This indicates that the model accurately captures the essential steps involved in sepsis patient care, and not being too much overfitted.

**Precision**, reflecting the model's ability to correctly identify patterns, attains a 93.19%. This suggests that the model effectively distinguishes between valid and invalid process executions.

**Generalization**, assessing the model's ability to handle new, unseen data, achieves a 96.8%. This implies that the model can generalize beyond the training data and adapt to real-world scenarios.

**Simplicity**, evaluating the model's ease of understanding and interpretation, reaches a 79.31%. This indicates that the model is not overly complex and can be readily grasped by hospital's managers.

In contrast, and as can be seen in the examples in the table, models discovered from a lower number of traces produced excessively high precision, which doesn't really allow the model to incorporate behaviors that were not observed in building the model (for example, in the above table, in the model discovered from the most common trace, the precision measure is 1, which indicates that behaviors that were not observed will not be possible according to this model). On the other hand, models built from more traces were too overfit, and their simplicity measure was also low, indicating overly complicated models (as can be seen, for example, in the model discovered from the five most common traces).

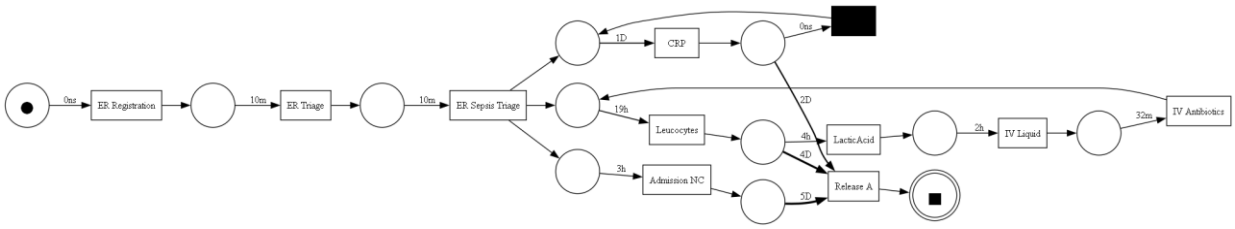
Therefore, the model with three variants strikes the optimal balance between performance and simplicity, making it the most suitable choice for representing the process flow. it does a good job of matching real data, is accurate, can handle new situations, and is easy to understand.

Now we will see the second and main goal of the conformance checking that we referred to above, with this goal we mapped the inconsistencies and identified the parts that are less good or appropriate in the process to identify opportunities for improvement of the process.

To perform this step, we examined the filtered log obtained manually, and we identified that many of the treated processes did not pass the initial stage of the checks (triage). This stage is essential and important for the patient, since it is in this stage that additional problems that may arise for him can be identified, so our recommendation is to pay attention that the patients do undergo all the necessary tests.

**Analysis of the model and insights into the processes**

The Temporal Information Plot evaluate if the process models aligned with medical guidelines by visualizing the timing of critical actions. This analysis aimed to determine if prescribed treatments and tests were administered within recommended timeframes.



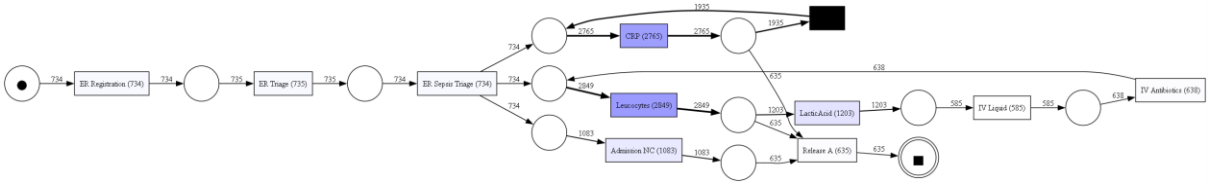

From an initial look at the data received, it can be concluded that there are events that take too long, but further investigation revealed that the cause of the delay was repeating the same test several times, and not a real failure in the process. For example, the guidelines state that patients should receive Antibiotics within 60 minutes of ER Sepsis Triage, followed by Lactic Acid tests within 180 minutes post-triage. However, a quick look at the visualization suggests that the average times don't align with these recommendations. Notably, the LacticAcid test appears to be conducted nearly 24 hours later. This delay is likely due to repeated testing, as our manual review of the logs indicates that severe violations of medical guidelines are not common.

We also looked at the analysis of the resources in the patient care processes. We mapped the different groups that perform the different processes, and tried to identify bottlenecks that would show us how to improve our model and the treatment of patients as we learned in the course. In the table shown below you can see the mapping between the different groups (marked with letters) and the different activities. In addition, we released a graph that analyzes the dependencies of the different groups and the different activities, and how many times each group performs each activity (this graph is in the appendices). The analysis of the resources shows that no significant bottlenecks were identified in the patient care processes, which indicates efficient utilization of the hospital's personnel.

| concept:name | |
|---|---|
| Admission IC | J, P, K, W |
| Admission NC | D, F, G, H, I, J, K, M, N, O, Q, R, S, T, U, V, P, X, W, Y |
| CRP | B |
| ER Registration | A, L |
| ER Sepsis Triage | A, L |
| ER Triage | C |
| IV Antibiotics | A, L |
| IV Liquid | A, L |
| LacticAcid | B |
| Leucocytes | B |
| Release A | E |
| Release B | E |
| Release C | E |
| Release D | E |
| Release E | E |
| Return ER | ? |

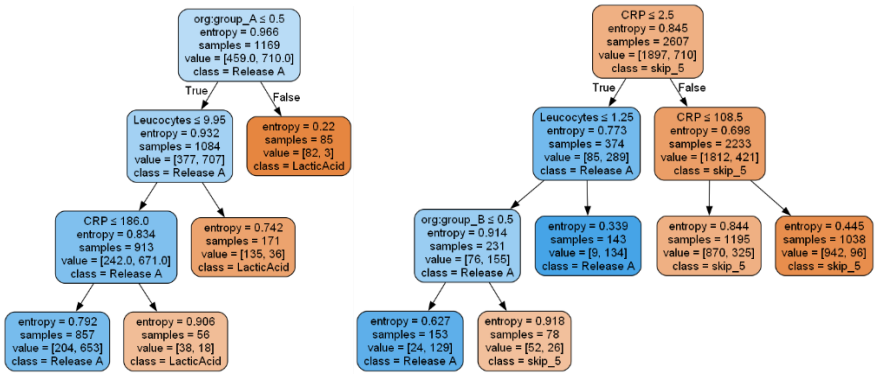
## Proposing improvements to the process based on methodologies learnt

We created Frequency Information Plot to provide insights into the occurrence rates of different activities in the process model. This plot helps us understand the distribution of activities and their relative importance in the overall process. We did this analysis with TBR (Token Based Replay).




We wanted to check if there is an event that seems unusual - an event that repeats less often, but as we can see that all the events are relatively common and therefore, we did not find any exceptions that can be removed.

we created classifiers for the decision points in our model. There are two decision points of interest: one after CRP, and another one after Leucocytes.




Making decisions based solely on these decision trees proves to be challenging. CRP and Leucocytes are important factors, but the trees show that there's a lot of uncertainty in the outcomes. This means that using only these two things might not be enough to make the right choice every time. We might need to consider other things or use more advanced methods to make better decisions.

## Summary

This research explores the process of treating sepsis patients through multiple analytical steps. We began by comprehending and cleaning the database, then proceeded to model discovery and conformance checking. We selected the optimal model, evaluated patient processes, and identified potential improvements in organizationasl procedures. The study concludes with efforts to enhance the model using additional techniques, including patient process time analysis and decision point mining, to gain deeper insights into sepsis treatment protocols.

The model we discovered could form a basis for more advanced research, which will help in effective treatment processes for sepsis patients.

**References**

[1] F. Mannhardt and D. Blinde. “Analyzing the trajectories of patients with sepsis using process mining”. English. In: RADAR+EMISA 2017, Essen, Germany, June 12-13, 2017. CEUR Workshop Proceedings. RADAR + EMISA 2017 ; Conference date: 12- 06-2017 Through 13-06-2017. CEUR-WS.org, 2017, pp. 72–80.

[2] Van Der Aalst, W. (2016). Process mining: Data science in action. Springer, Berlin, Heidelberg. Pp. 167-177

[3] Van Der Aalst, W. (2016). Process mining: Data science in action. Springer, Berlin, Heidelberg. Pp. 222-235

[4] Van Der Aalst, W. (2016). Process mining: Data science in action. Springer, Berlin, Heidelberg. Pp. 243-274

[5] Van Der Aalst, W. (2016). Process mining: Data science in action. Springer, Berlin, Heidelberg. Pp. 294-296

## Appendices

A link to our code:

https://github.com/Noam-Diamant/Process-Mining/blob/main/sepsis_project.ipynb

Description of the database and preprocessing the database

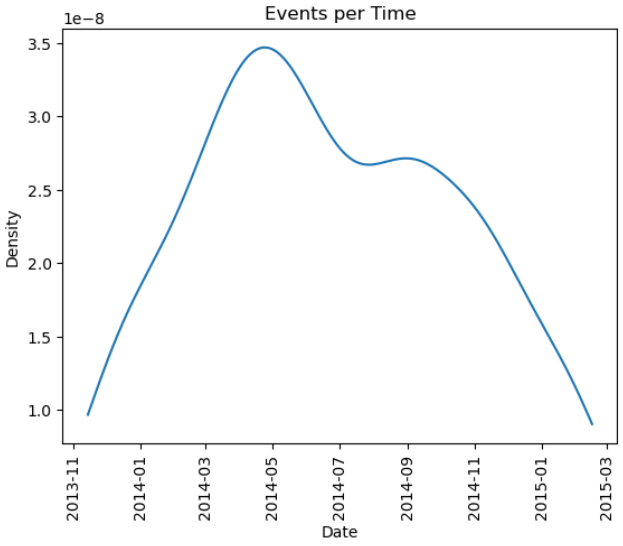
The most common variants of the database before the filtering we applied:

| | variant | count | percentile |
|---|---|---|---|
| 21 | (ER Registration, ER Triage, ER Sepsis Triage) | 35 | 3.333333 |
| 24 | (ER Registration, ER Triage, ER Sepsis Triage, Leucocytes, CRP) | 24 | 2.285714 |
| 95 | (ER Registration, ER Triage, ER Sepsis Triage, CRP, Leucocytes) | 22 | 2.095238 |
| 3 | (ER Registration, ER Triage, ER Sepsis Triage, CRP, LacticAcid, Leucocytes, IV Liquid, IV Antibiotics) | 13 | 1.238095 |
| 36 | (ER Registration, ER Triage, ER Sepsis Triage, Leucocytes, CRP, LacticAcid) | 11 | 1.047619 |

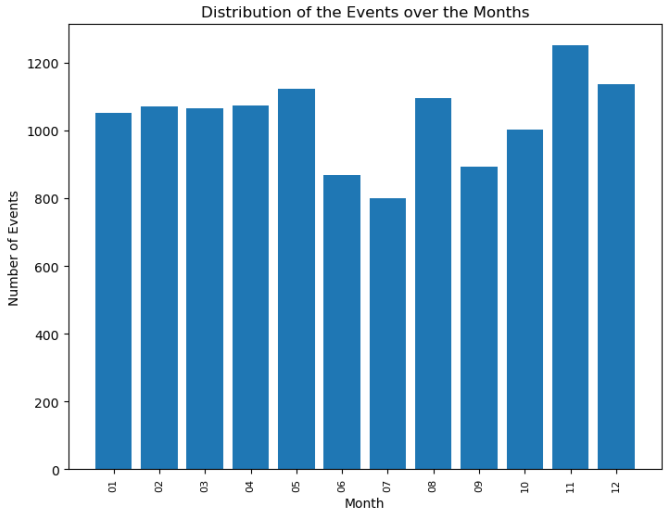
The most common variants of the database after the filtering we applied:

| | variant | count | percentile |
|---|---|---|---|
| 122 | (ER Registration, ER Triage, ER Sepsis Triage, Leucocytes, CRP, Admission NC, Release A) | 5 | 0.681199 |
| 183 | (ER Registration, ER Triage, ER Sepsis Triage, LacticAcid, Leucocytes, CRP, IV Liquid, IV Antibiotics, Admission NC, Release A) | 4 | 0.544959 |
| 48 | (ER Registration, ER Triage, ER Sepsis Triage, CRP, Leucocytes, Admission NC, Release A) | 4 | 0.544959 |
| 213 | (ER Registration, ER Triage, ER Sepsis Triage, Leucocytes, LacticAcid, CRP, IV Liquid, IV Antibiotics, Admission NC, CRP, Leucocytes, Release A) | 4 | 0.544959 |
| 269 | (ER Registration, ER Triage, ER Sepsis Triage, Leucocytes, CRP, LacticAcid, IV Liquid, IV Antibiotics, Admission NC, Leucocytes, CRP, Release A) | 4 | 0.544959 |

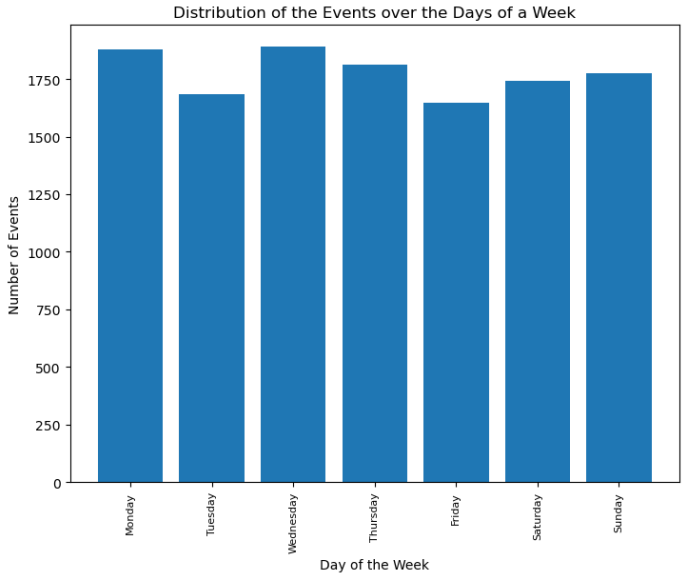
Distribution of events by date:



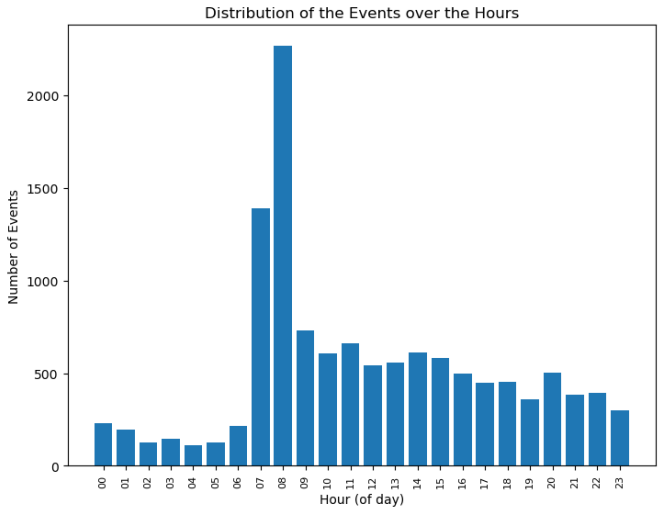

Distribution of events by month:




Distribution of events by day:




Distribution of events by hour:



Process Discovery

Approvals of the lecturer:

Here we asked him about the different approaches we made:

23 ביולי 2024 בשעה 15:58 **נועם דיאמנט** <noamd1234@gmail.com>
אל: Izack Cohen <izack.cohen@biu.ac.il>

אוקיי, אבל לגבי היחס טריין טסט שיצא לנו זה גם בסדר?
יום טוב,
נועם דיאמנט.

בתאריך יום ג', 23 ביולי 2024 ב-15:53 מאת Izack Cohen <izack.cohen@biu.ac.il>:
[טקסט מצוטט מוסתר]

---

23 ביולי 2024 בשעה 16:54 **Izack Cohen** <izack.cohen@biu.ac.il>
אל: "noamd1234@gmail.com" <noamd1234@gmail.com>

מה שחשוב הם התוצאות.

Here we asked him about our train – test ratio:

23 ביולי 2024 בשעה 15:58 **נועם דיאמנט** <noamd1234@gmail.com>
אל: Izack Cohen <izack.cohen@biu.ac.il>

אוקיי, אבל לגבי היחס טריין טסט שיצא לנו זה גם בסדר?
יום טוב,
נועם דיאמנט.

בתאריך יום ג', 23 ביולי 2024 ב-15:53 מאת Izack Cohen <izack.cohen@biu.ac.il>:
[טקסט מצוטט מוסתר]

---

23 ביולי 2024 בשעה 16:54 **Izack Cohen** <izack.cohen@biu.ac.il>
אל: "noamd1234@gmail.com" <noamd1234@gmail.com>

מה שחשוב הם התוצאות.

Analysis of the model and insights into the processes

Social network graph: