Sepsis
Event log analysis
Process Mining 2024

Kacper Dobek 148247

May 8, 2024

Instructor: Tomasz Pawlak, PhD, DSc

1 Description of the problem

This report presents the analysis of the handling of patients suspected of sepsis in one of the Dutch hospitals. Sepsis is the body's extreme response to an infection and is potentially life-threatening. According to the authors of the original project [1], process mining can provide vast insights into the trajectory of patients in a hospital and verify whether the daily clinical practice follows medical guidelines.
The analyzed event log [2] contains 1050 real-life traces recorded by the hospital ERP system over the course of 18 months. More precisely, there are 15,214 events and 16 different activities. Furthermore, 39 data attributes are stored, for example, the results of medical trials. The median case duration is 5.3 days, and the mean case duration is 28.5 days. Cases were registered between 07.11.2013 and 05.06.2015.
The code and other resources used for the preparation of this report are available at `https://github.com/kapiblue/sepsis-event-log-analysis.git`.

2 Preliminary analysis of the event log

The preliminary analysis presented in this section is performed on the full event log. We can get an idea of a typical process flow by looking at the process map shown in Figure 1. The process contains loops, with the loop between Leucocytes and the C-Reactive Protein (CRP) being the most prominent example.

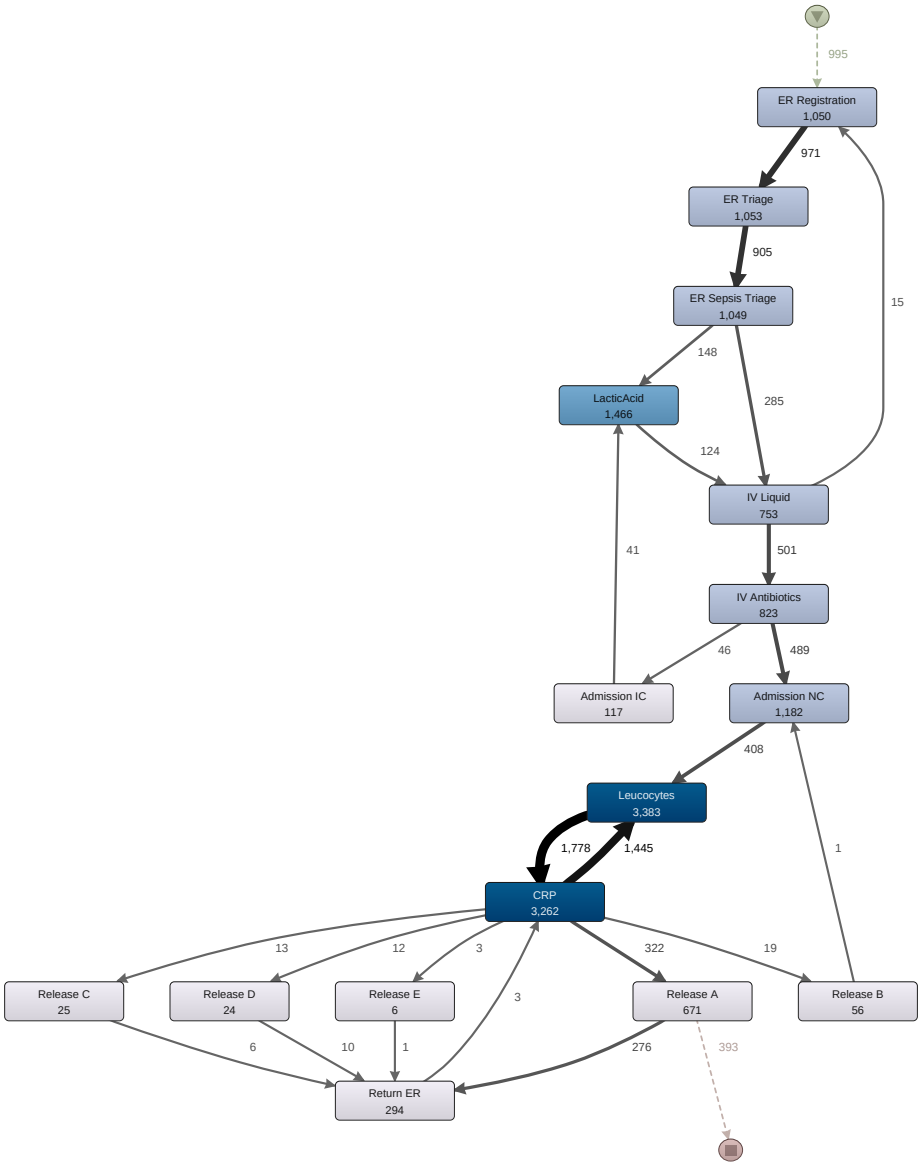


Figure 1: Process map generated from the entire log using Disco. A darker color indicates higher absolute frequency.

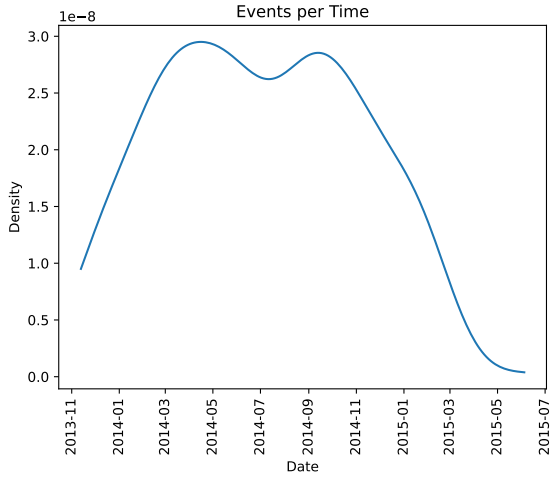
Additionally, we look at five of the most frequent process variants. Interestingly, while all begin with the Emergency Room (ER) registration, none of them finish with a discharge. This is a problem that needs further inspection and will be addressed at the end of this section.

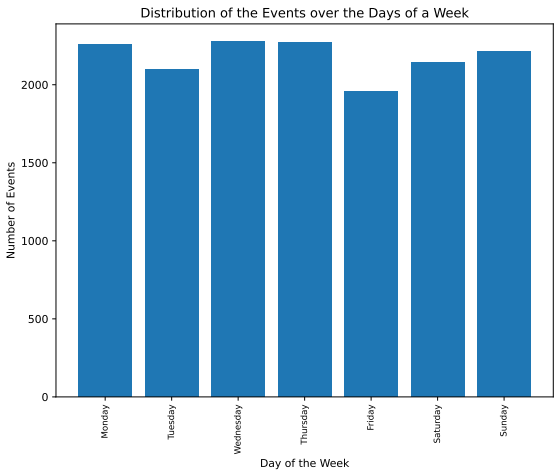
variant	count	percentile
(ER Registration, ER Triage, ER Sepsis Triage)	35	3.33
(ER Registration, ER Triage, ER Sepsis Triage, Leucocytes, CRP)	24	2.29
(ER Registration, ER Triage, ER Sepsis Triage, CRP, Leucocytes)	22	2.10
(ER Registration, ER Triage, ER Sepsis Triage, CRP, LacticAcid, Leucocytes, IV Liquid, IV Antibiotics)	13	1.24
(ER Registration, ER Triage, ER Sepsis Triage, Leucocytes, CRP, LacticAcid)	11	1.05

The relative frequency of activities is given in the table below. There are two relatively frequent activities – Leucocytes and CRP. In the column to the right, we can see which group performs the activity.

Activity	Relative frequency (%)	org:group
Leucocytes	22.24	B
CRP	21.44	B
LacticAcid	9.64	D, F, G, H, I, J, K, M, N, O, Q, R, S, T, U, V, P, X, W, Y
Admission NC	7.77	C
ER Triage	6.92	A, L
ER Registration	6.90	A, L
ER Sepsis Triage	6.89	A, L
IV Antibiotics	5.41	A, L
IV Liquid	4.95	E
Release A	4.41	?
Return ER	1.93	J, P, K, W
Admission IC	0.77	E
Release B	0.37	E
Release C	0.16	E
Release D	0.16	E
Release E	0.04	

Next, we examine the distribution of events over time. The chart to the left shows the density plot. Surprisingly, we observe a somewhat bimodal distribution with modes in the spring and fall of 2014. The second plot depicts the distribution over the days of the week. There is no drastic variation here. The fewest events are registered on Friday.

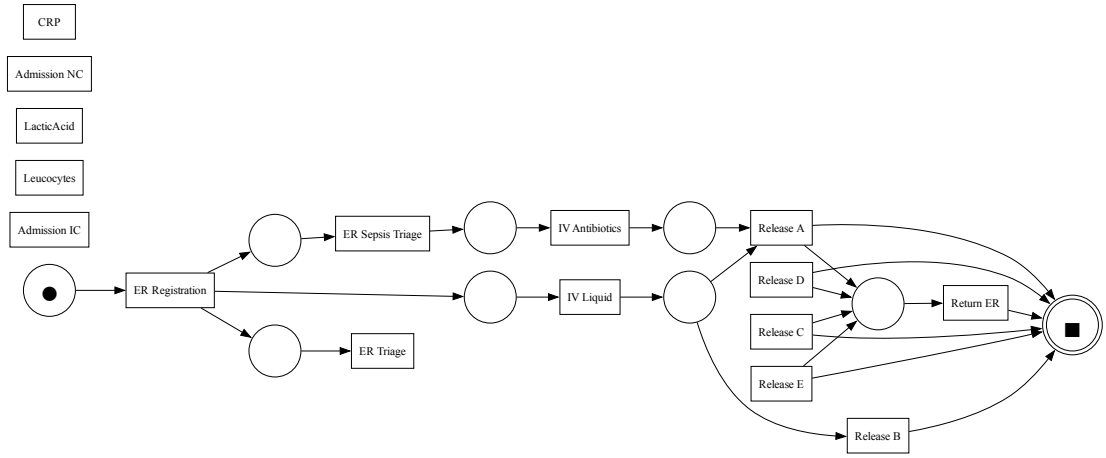




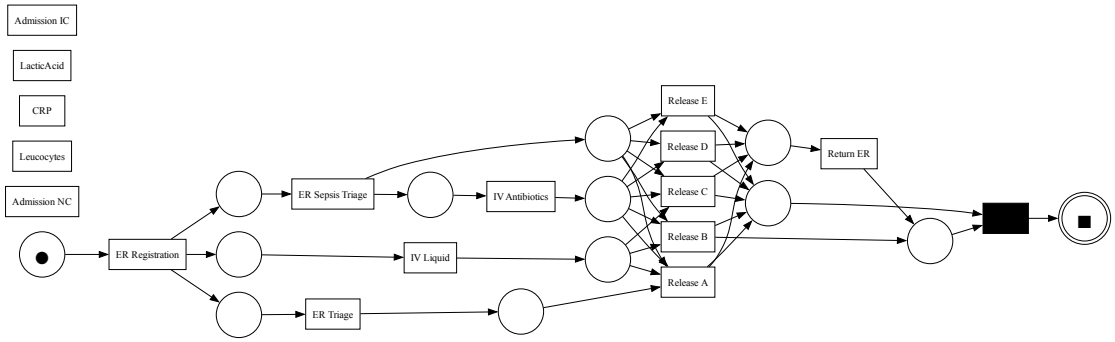
Our event log contains some incomplete cases. A patient's trajectory in the hospital is expected to begin with registration and finish with some form of discharge. Hence, from this time on, the analysis will concern the filtered event log.

3 PM4Py analysis

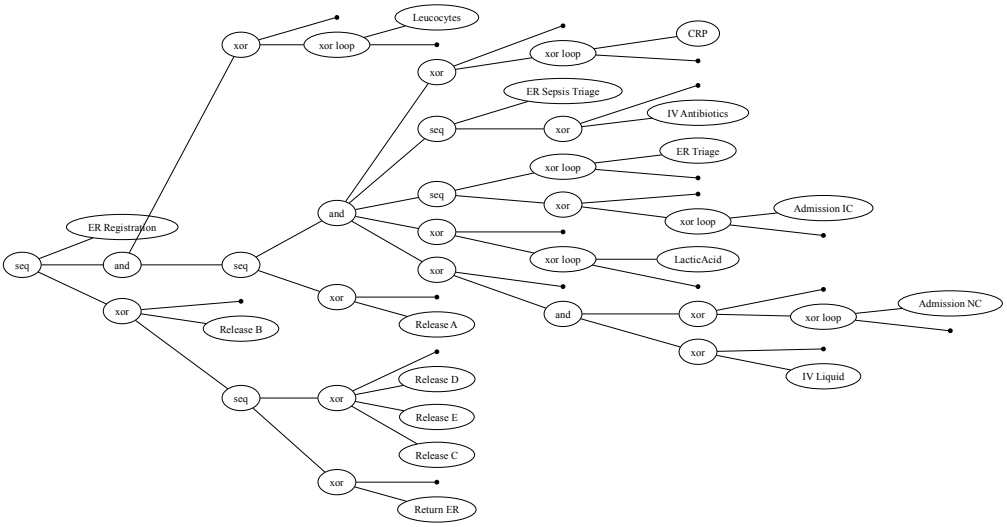
For the purpose of process model discovery using a Python package PM4Py, the event log was filtered using the start and end activities. The valid start activity is *ER Registration*, and the valid end activities are *Release A, Release B, Release C, Release D, Release E, Return ER*. The filtered event log contains 734 traces.
Let's build a simple Petri net using the whole event log:



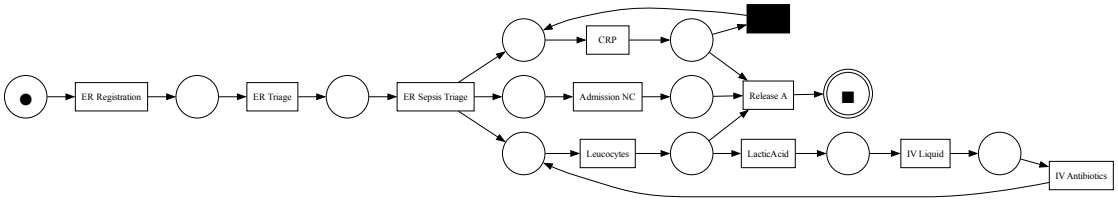
There are many deficiencies in the above-presented model, including the lack of soundness. Moreover, the basic α algorithm cannot handle short loops.
As an improvement, we build the process model using the $\alpha+$ algorithm. Unfortunately, this model is not sound either.



Again, we observe disconnected activities and the model is not sound. The PM4Py library allows us to build other models apart from Petri nets. Therefore, we build a process tree – a model which is sound by construction. This results in a quite convoluted model:



Even though the process tree representation has its advantages, let's return to the Petri net representation, as it is easier to analyze. For the final improvement, the model is built using the top three process variants and the inductive algorithm. This results in a sound model, that contains no dead parts. The final state is reachable from each of the states.

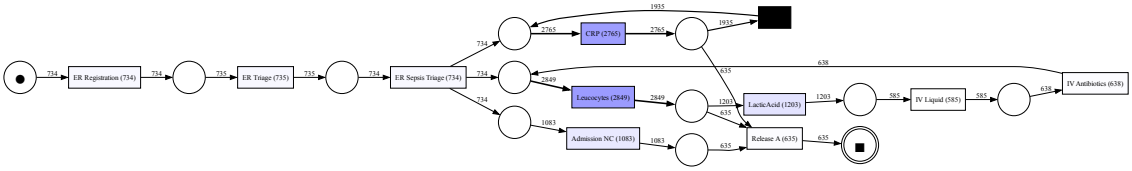


In fact, the model looks much different from the one shown in Figure 1. We do not have a loop between Leucocytes and CRP, however, these activities have their separate loops. The Leucocytes loop contains other activities – LacticAid, intravenous therapy (IV) Liquid, and IV Antibiotics. It is quite surprising, as the Leucocytes and CRP are often interleaved in the event log. Hence, one of the most common deviations from the typical flow would be a sequence of these two activities. Apart from that, sometimes the flow in the beginning is violated. For instance, we may observe some medical tests done before ER Sepsis Triage.

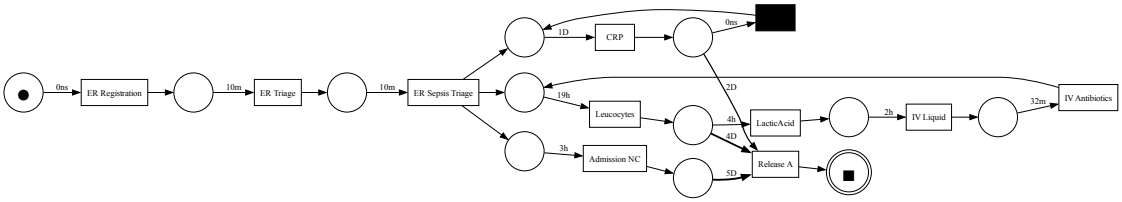
Naturally, this is a rather simple model. Building the model using more than three most frequent variants increases the average fitness, but significantly decreases precision and simplicity. Hence, the final model was built using only the three most frequent variants. The metrics for the model are presented in the table below.

Average Fitness	Precision	Generalization	Simplicity
0.77	0.93	0.97	0.79

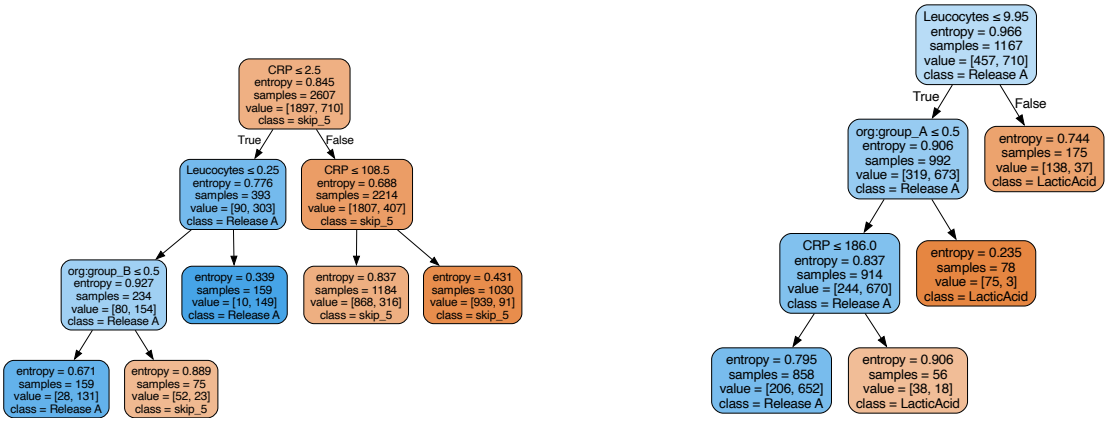
In the following visualizations, we see the log replayed with Token-Based Replay (TBR). The plot below shows the model with the frequency information.



Similarly, we plot the temporal information. It allows us to check the process model against medical guidelines [3]. For instance, patients should be administered Antibiotics within one hour after ER Sepsis Triage. Next, LacticAcid measurements should be performed within three hours after the triage. Just a glimpse at the plot reveals that the mean times do not indicate that the guidelines are met. Indeed, it takes almost 24 hours before the LacticAcid is performed. This is likely due to cases when the tests are redone, as the manual log analysis shows that the medical guidelines are not severely violated. The temporal analysis can be as well performed on the process maps generated by Disco.

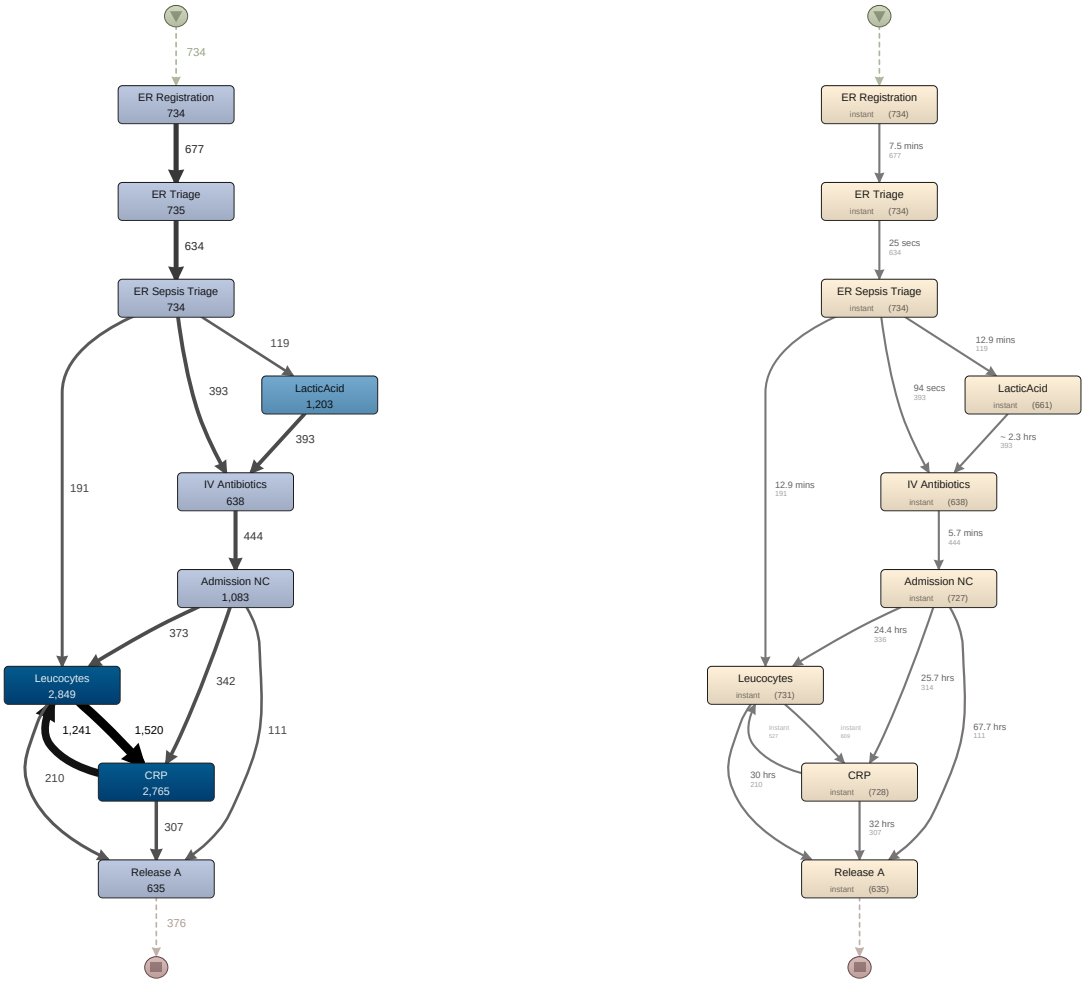


Finally, we can create classifiers for the decision points in our model. There are two decision points of interest: one after CRP, and another one after Leucocytes. One should note that the leaves are impure, which may indicate that the decision is complex, or the model is oversimplified.



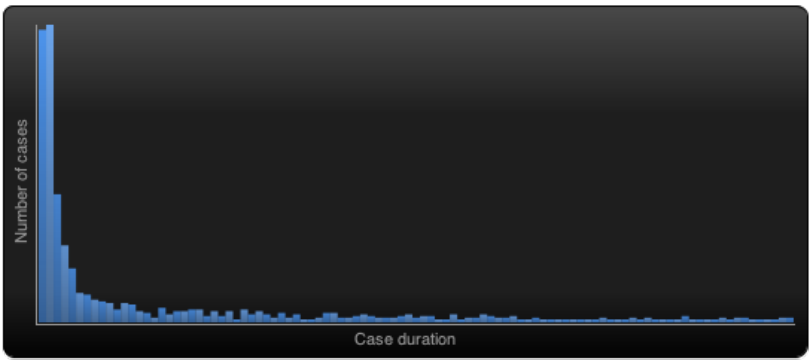
4 Disco analysis

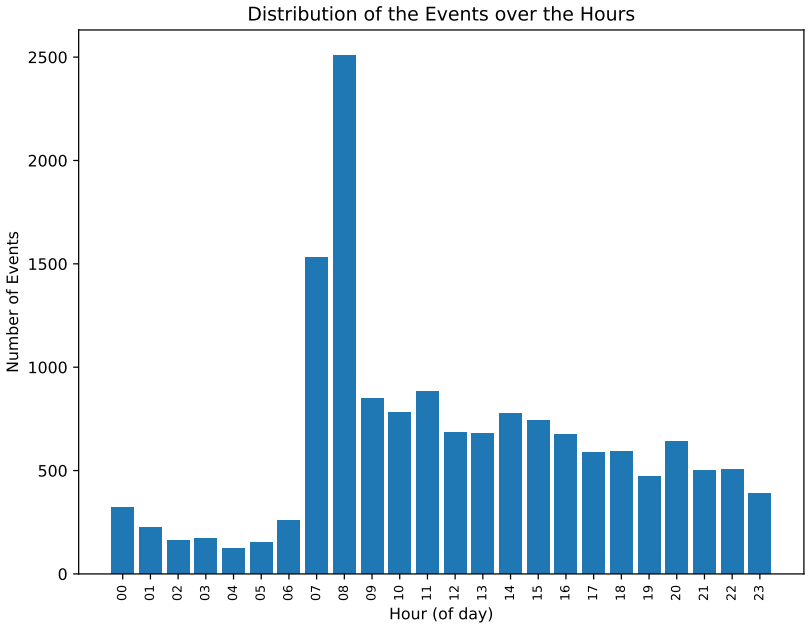
This section provides a similar analysis, using another process mining tool called Disco. The event log is filtered in the same way as described at the beginning of Sec. 3.
Below, we present a process map discovered using the filtered event log. Since the process is rather complex, the level of detail seen in the map was adjusted by setting the Activity slider at 20% and the Paths slider at 5%. The process map to the left contains the information on the absolute frequency, while the one to the right shows median duration and case frequency as the second criterion.



As expected, the process begins with registration, then, there is the ER Triage and ER Sepsis Triage. We observe a loop between Leucocytes and CRP – the two most frequent activities. The process ends with the most common form of discharge – Release A. All activities are recorded as instant. Next, activities at the beginning of the process are executed in a short time (median below 13 minutes). This is expected when dealing with a life-threatening situation. The longer waiting times appear towards the end of the process

when the patients are waiting to be discharged. Interestingly, some activities are looped. For example, it is quite common for patients to have their Leucocytes measured two or three times in a row over several hours. The same applies to LacticAcid. Presumably, the goal was to observe the change or to correct the measurements. As for the longest-running cases, they concern patients, who returned after the release. Precisely, the longest case ran for 1 year and 57 days. The patient spent 5 days at the hospital and returned after 1 year and 52 days. The plot below exported from Disco shows that most cases are relatively short, so the described example is clearly an outlier.





Finally, we look at the periodicity of activities aided by a visualization from PM4Py. Most activities happen at 7 and 8 AM. The further analysis in Disco reveals that this is when measurements are performed most of the time (CRP, Leucocytes).

5 Conclusion

On balance, the exploration of the sepsis event log provided many interesting results. We explored the capabilities of two tools, the Python PM4Py library, and Disco. It is hard to say which tool is better, as one of them offers a programmatic approach, and the other provides a pleasant GUI. Speaking about the discovered process models, it seems like Disco creates more elegant models out of the box. However, the tool is limited when we would like to calculate metrics. On the other hand, it would be desirable to have more developed documentation for PM4Py. One useful thing I discovered while working with the library is the possibility of saving plots as PDF files.

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] Felix Mannhardt. *Sepsis Cases - Event Log*. 2016. DOI: 10.4121/uuid:915d2bfb-7e84-49ad-a286-dc35f063a460. URL: https://data.4tu.nl/articles/dataset/Sepsis_Cases_-_Event_Log/12707639/1.

[3] Andrew Rhodes et al. "Surviving Sepsis Campaign: International Guidelines for Management of Sepsis and Septic Shock: 2016". In: *Intensive Care Medicine* 43.3 (Jan. 2017), pp. 304–377. ISSN: 1432-1238. DOI: 10.1007/s00134-017-4683-6. URL: http://dx.doi.org/10.1007/s00134-017-4683-6.