Milestone 3: Process book

Léonard Vaney, Hojjat Karami, Jeremy Dezalos

June 4, 2023

1 Introduction

1.1 Goal

Our main goal was to provide an interesting and appealing presentation for novice of EHR data on a medical environment. To accomplish that we based our work on the lectures of the course and different reference visualizations that we describe on a dedicated subsection.

1.2 MIMIC dataset

Our project has been built around the MIMIC-IV dataset which contains deidentified electronic health records (or EHR). It contains various kind of data that we had to deal with such as time events, IDs, codes, and complex dependencies between the different files of the dataset. We first decided to keep only a small subset of patients containing at least one ICD code (diagnostic) with the word "sepsis" in it. It reduced our original dataset from 60'000 to less than 8000 patients. Sepsis is a life-threatening condition which leads us to have more interesting data with respect to ICU transfers, treatments, length of stay, outcome at the end of hospitalization, the variety of additional ICD codes. Thanks to this filtering it is easier to compare different treatments for different patients and their evolution over time. For our two mains visualization we extracted the value of organic elements found in patients over time to be able to visualize the impact of the illness and their respective treatments which has been stored in the resources folder in three different files: dict_map_states.json, test.json and tsne_datavis.json .

2 Exploration phase

2.1 Data analysis

Our data analysis revolved around information found on patients that at some points were diagnosed with sepsis. First we were able to confirm the lethality of the illness by looking at the outcome of stay from figure 1.

We also explored the different ICD codes assigned to patients and the top 15 of those codes shows a variety of different diagnostics that can be found on figure 2.

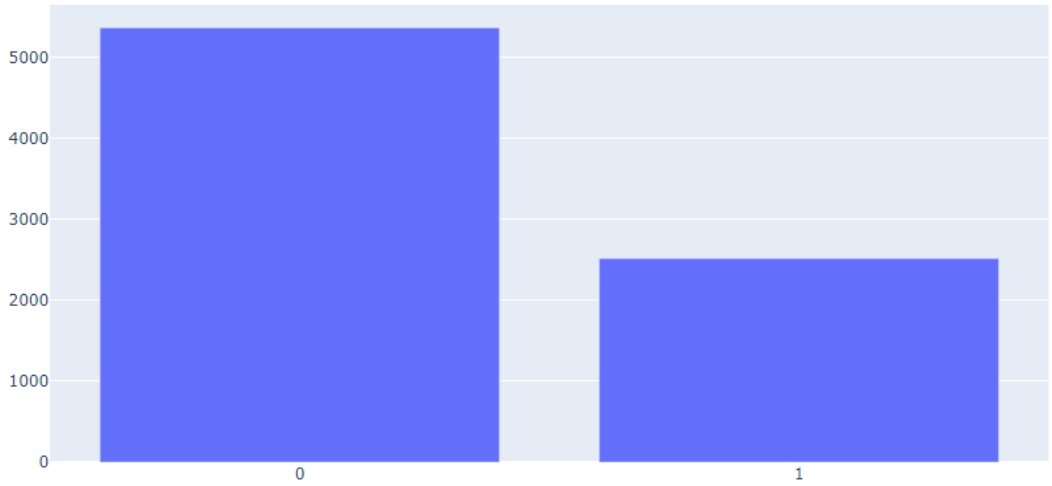


Figure 1: Outcome of stay: 1 is death, 0 is alive

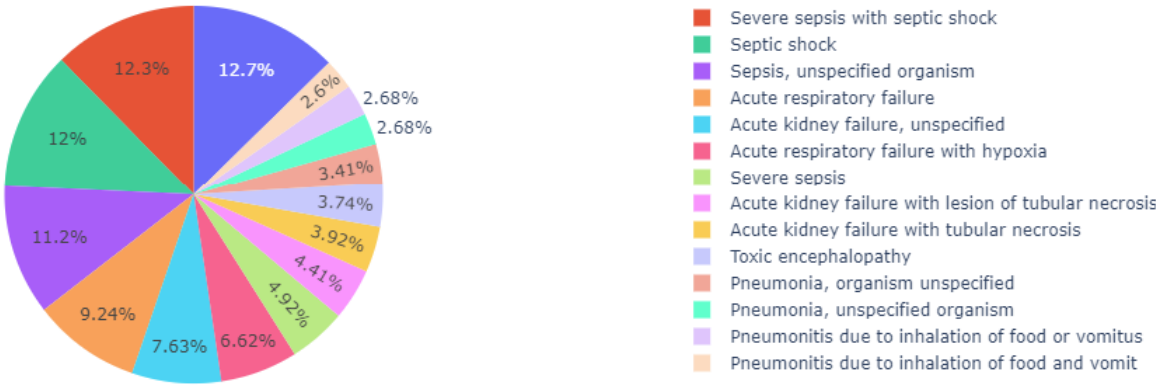


Figure 2: Pie chart of the top 15 ICD codes

More graphs and results have been discussed in milestone 1 and all the code related to data analysis can be found in the following files: DataExploration.ipynb, hk_data_exp.ipynb and SomePlots.ipynb .

2.2 Draft of visualizations

During our brainstorming sessions we thought about how to build our website and visualization while keeping the accessibility and readability in mind. Thus we arrive at the conclusion that a separation of the website in multiple panels would be an interesting idea to keep things simple and convey useful insights on the data without users being overwhelmed by too many options. You can find a sketch of this on figure 8.

It contains different panels with explanations on how the user would be able to navigate through the visualizations. It includes a panel for the dataset explanation, for time series of the ward transfer and for event alignment. However we didn't keep the time series panel for the final version because we tried to focus on the patients data themselves instead of the transfer of patients. The event alignment panel has been moved to the potential extension as it would require a lot of work to have a functional

and aesthetically pleasant result.

2.3 References

Most of our references comes either from the lectures or the D3 gallery. We used the scatter plot template from the D3 gallery as it is a simple solution for our needs.

For the more complex and optional visualizations, we looked at a paper on the state-of-the-art EHR visualization called EHR STAR. Our main inspiration for the extensions are based on some papers linked in section 4 such as OutFlow (Figure 4) for a Sankey diagram of temporal events and EventFlow for the alignment of temporal events on one point.

3 Implementation

3.1 Website

The final version of the website has been built around a template that can be found here. For the website, we wanted to keep everything to a single web page to make the navigation fluid. However we didn't to have everything cluttered at the same place so we decided to organize the page in *panels*. Each panel would contain a single bit of information to make things understandable for the user. More specifically, each panel contains only a brief description of the project or a single visualization or multiple visualizations that are similar in terms of data displayed.

3.2 Visualization

3.2.1 Multi-patient view

The multi-patient view is a way for the user to quickly select a patient to access its data that has been collected during their stay in the ICU. The raw data is however not clearly readable and we had to make it understandable for the user. By using t-Distributed Stochastic Neighbor Embedding (or t-SNE), a machine learning technique that preserves local relationships, we were able to map the data a lower dimension so that it would be displayable on a scatter plot. On this chart, each data point corresponds to a patient who have been mapped with t-SNE into a lower dimension. Each patient also received a prediction from our model determining if the patient had a sepsis or not. Predictions are represented with a color and false predictions have been highlighted with a red circle. We designed this view to help the user understand how some factors of the patient's stay are more relevant for the diagnosis.

When clicked on, the window automatically scrolls down to the single-patient view for the selected patient.

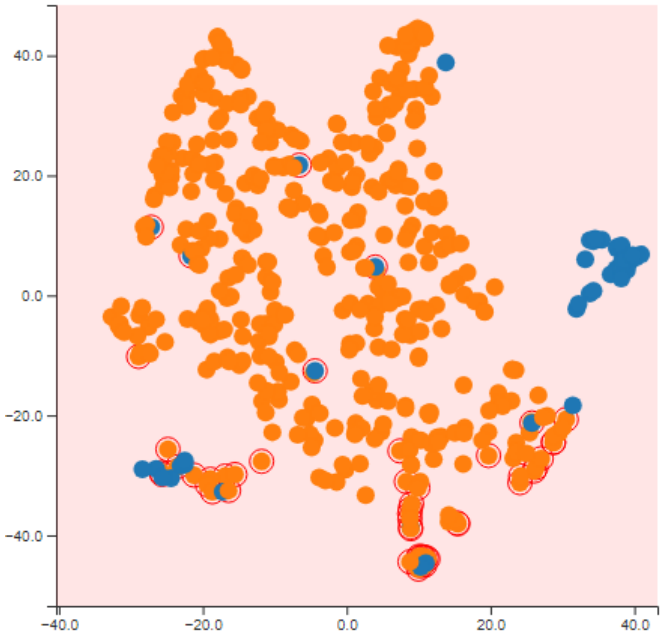


Figure 3: Final multi-patient patient view

3.2.2 Single-patient view

The single-patient view is a scatter plot of time series containing health record of a patient at the ICU at irregular intervals. The weights that our model found the most relevant for its prediction are highlighted in green. The saturation of the highlight depends on how much a measurement was relevant.

The panel didn't changed much from our first design as you can see in the figure below. The principal flaw of our original idea is that we wanted that the measurements labels would re-arrange themselves from the value with the highest standard deviation value to the lowest at the timestamp that is hovered by the mouse. The consequence of this choice is that the visualization was erratic because it changed whenever the mouse moved on the screen. Also the user needed to be really precise when selecting the timestamp to reorder. To fix this issue, we decided that the input needed from the user would be a click on any measurement taken at the desired time to sort the different measurements.

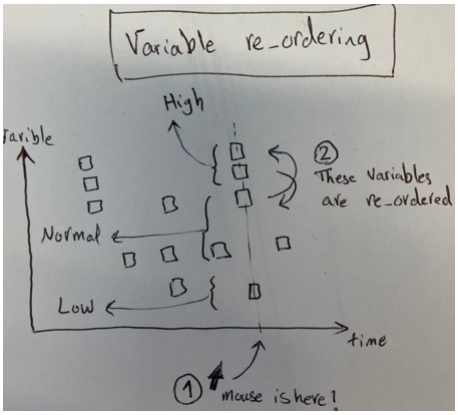
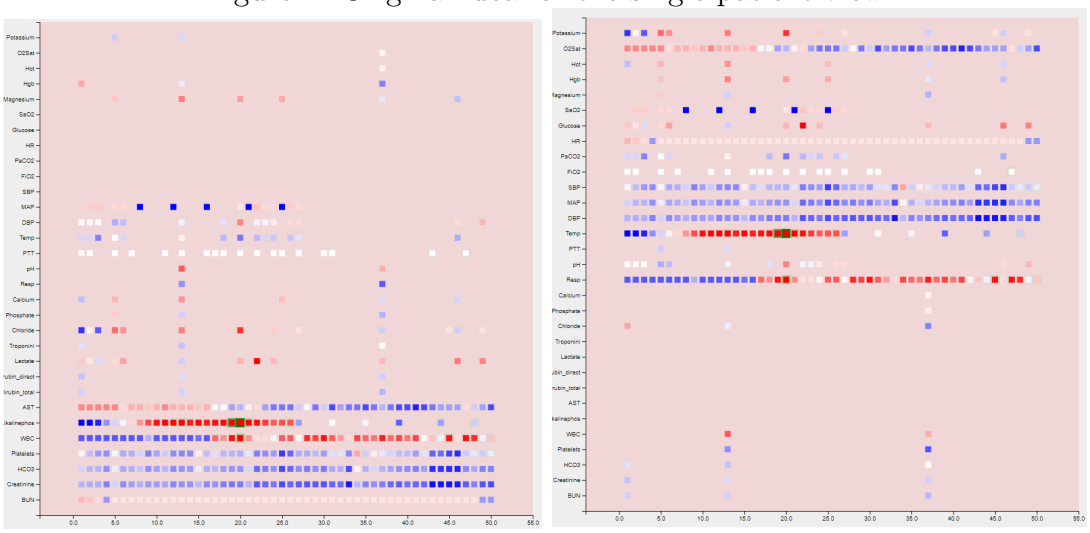


Figure 4: Original idea for the single patient view



Figures 5 and 6: final view of the single-patient view

The next point we needed to address is that the time series contained a lot of timestamps, so much that it became hard to see to which type a measurement a data point corresponded to. We fixed this issue by slightly darkening the zone around the data points that corresponded to where the mouse was located to make the visualization clearer for the user.

We also added the possibility for the user to filter the measurements by setting the minimum standard deviation a data point must have to be displayed on screen. This new feature was introduced to help the user understand what was wrong with the patient because a measure with high deviation from the norm is almost never a good sign for the health of a patient.

3.2.3 Additional single-patient view

After completing the single-patient view mentioned previously, we realized that the current visualization lacked visual cues telling the user when the patient's vital signals were concerning. This is why we introduced a new way to explore data from a single patient.

The new chart is a stack area chart that gives the opportunity to see quickly when the status of a patient is problematic. The time series is transformed to compute by how much all the measurement derive from the norm. We also added the possibility to filter the data points by a minimum standard deviation.

To give a sense of continuity with the previous visualization, when hovering a timestamp the data is darkened in both charts. Also, modifying a chart (zooming or panning) applies the same modification to the other chart.

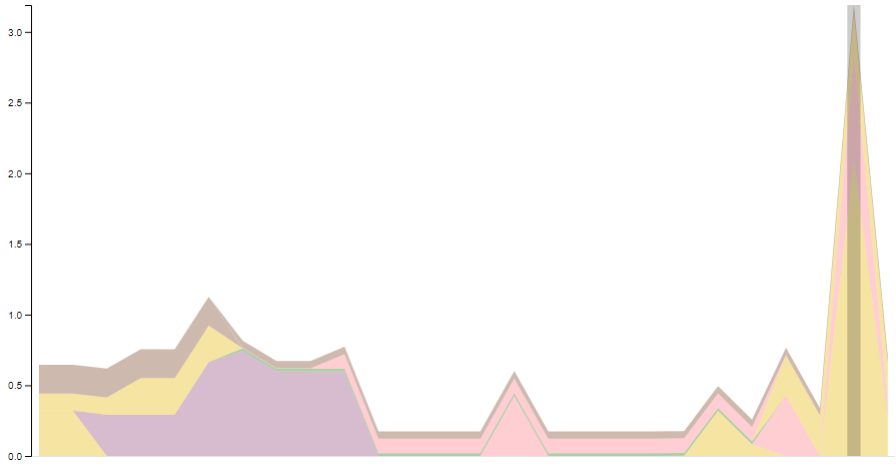


Figure 7: New panel showing total absolute deviation of the measurements

4 Peer assessment

The group met each week to discuss what each member did during the last week, brainstorm ideas for the upcoming deadline, work on the project and assigned to each member what he must implement by next week. Here is the breakdown of the parts of the project completed by each member for every milestone:

4.1 Milestone 1

Hojjat: Data set explanation + related work
Jeremy: Exploratory data analysis
Léonard: Problematic

4.2 Milestone 2

Hojjat: Design of the visualizations
Jeremy: Skeleton of the website
Léonard: Stack bar showing patients moving between hospital wards (not kept in final website)

4.3 Milestone 3

Hojjat: Website + visualizations
Jeremy: Variable re-ordering in the single-patient view + report (part 3) + screencast montage
Léonard: Report (part 1 and 2) + old first draft of website (not kept at the end) + screencast

5 Appendix

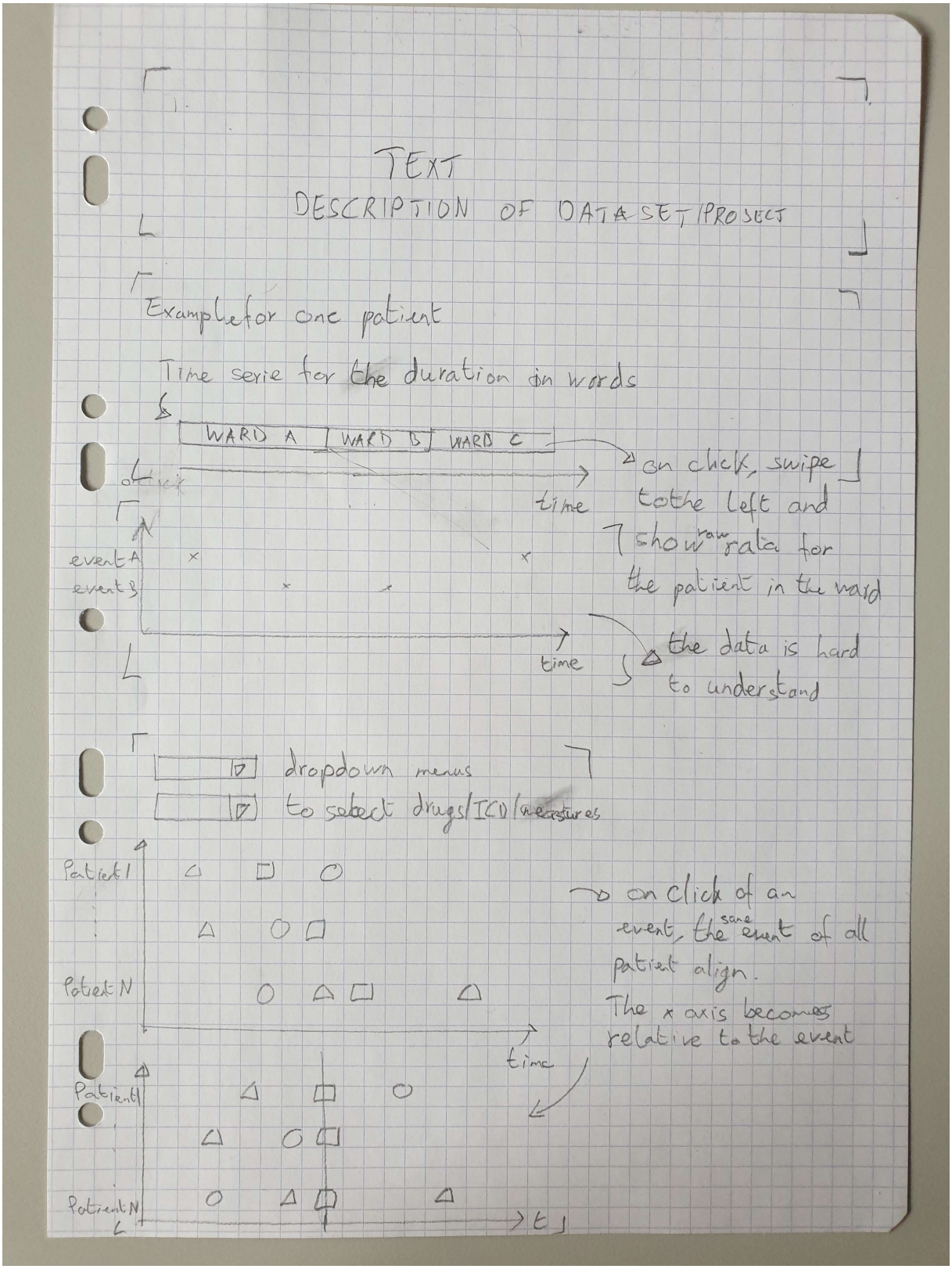


Figure 8: Sketch of the early design of the website