Interdisciplinary Project Report
Feasibility study for a patient smartwatch data set in cancer research

Paul Erpenstein (12107369)

July 2023

Main Supervisor:
Anna Sophie Berghoff, PhD Assoc. Prof. Priv.-Doz. Dr. med.univ. et scient.med. Medical University of Vienna
Department of Medicine I - Division of Oncology

Co-Supervisor:
Allan Hanbury, Univ.Prof. Dr. Technical University of Vienna
Institute for Information Systems Engineering

Accompanying lecture: PS Epidemiological Methods 851.099
To be completed in WS 2023/24

Abstract Chemotherapy is an important treatment option for dealing with various types of cancer. The aim of this project was to assist a research unit at the general hospital of Vienna with carrying out a study concerning cancer patients undergoing systemic treatment. More than 50 patients were equipped with smartwatches and a huge data set about their daily activity and vital signs was captured. This report describes the research questions, process and insights gained from reorganizing and exploring this data set. This project assessed the viability of using this data set as a basis for medical research. While there are huge gaps in the recorded data, the measurements are consistent and of acceptable quality. This project included the implementation of an interactive data dashboard in order to visualize the restructured data. The dashboard enables the researchers to explore the data by themselves and gain an intuitive understanding for it. Taking the sparsity into consideration, a treatment-cycles-based data model was developed. This approach models only treatments cycles for which enough data is available. This will be used for modelling the probability of delays and treatment dropouts on a cycle-by-cycle basis.

1 Introduction

Systemic treatment including chemotherapy, targeted therapies and immunotherapy are important treatment options, particularly for patients with advanced cancer. Over the last decade, treatment efficacy has risen significantly. However, patients still frequently suffer from adverse effects. Especially, extended fatigue and malaise after intense systemic treatment. This frequently results in delays and reductions of the prescribed treatment. To avoid side effects and thereby secure the quality of life of patients, a more personalized “health adapted” treatment approach would be warranted. The goal of such an approach would be to adapt the treatment intensity based on the patient’s individual level of health, thereby preventing adverse effects.

In order to investigate the feasibility of the aforementioned approach, a pilot study including more than 50 patients with advanced cancer, undergoing systemic treatment, was conducted. These patients were equipped with smartwatches or so-called wearable devices (WAD). The WADs were used

to monitor vital signs and activity data. In this pilot study, we aim to investigate the feasibility of using WADs and the resulting data for assessing patient health.

The data collected includes daily steps taken, pulse, stress and beat-to-beat intervals of a patient. The volume of data collected is rather large, since there is very granular data, that stems from a time-window larger than 6 months. This time-series data set will be the focus of this project. The main focus of this project was data wrangling and exploration and is only a small fraction of the larger PhD project of Dr. Dominikus Huber.

The data were collected using Vivosmart 4 watches manufactured by Garmin. The WAD doesn't enable direct data access, and it can only be accessed by using a third-party service provider, over a cloud interface. When initially retrieving the data, it looked like they were too unstructured to be usable for the research. The goal of this project was to investigate a restructuring of the data and implement it if possible. In a more general sense it was about unlocking the information contained in the data and unlocking it for the medical researchers that collected it.

2 Research questions

When initially planning this project, the focus was much more on succinct questions about the data as well as the possibility of modelling. Very little attention was given to the overall goals that the project aimed to complete. Therefore, my adapted research questions are as follows:

1. **Basic data exploration:** What is the original structure of the data?
2. **Data cleaning and restructuring:** In what state should the data be in for it to be of use to the medical researchers? Are the collected measurements actually feasible?
3. **Requirements engineering:** What do the medical researchers need in a more general sense to attain the goals of the study? How can we overcome the sparsity of the data set?
4. **Data analysis:** Are there any noteworthy patterns in the data?

The main adaptation from the initial research questions is a shifting of the focus from very specific questions about the data, to a more general approach that deals with creating a baseline that actually makes this analysis possible. My previous project proposal dealt with questions like the following:

- Which data was actually acquired?
- Do patient wear the fitness tracker and if yes how long?
- Do any factors like age, type of cancer, type of therapy, gender, etc. impact the wear-time?
- Did the patients use the WADs on a regular basis, and does this result in a data set of acceptable size?

While all of these question were answered over the course of the project, they are not really the core of the project. The much more interesting part was that of requirements engineering in collaboration with the medical researchers. By this, I mean the process of finding out how to best assist the study team in attaining novel medical insights.

3 Process and Methodology

In this section, I start off with describing important aspects that characterized the process of this project. This creates a foundation for explaining the challenges faced and the process of overcoming them.

3.1 General structure of the process

The most important part of the project were regular meetings with Dr. Huber and Dr. Berghoff. These usually followed a similar pattern. In the beginning, we would share new developments concerning the project. I would often present the two doctors with new insights about the data in the form of visualizations or statistics. This would lead into a more general discussion about possible next steps. These meetings served as an important touchstone. As new insights about the data trickled in, we adapted our view on how the project goals could be achieved. This meant that we followed an iterative process, where we adapted the plan as new information became available to us.

As already mentioned, visualizations were an important part of communicating aspects about the data to the domain experts. To this end, an interactive visualization was developed, so the medical researchers could get a better sense of the data on their own.

Another important aspect of the process, was unit testing. Multiple, relatively complex software components had to be developed in order to restructure and subsequently visualize the data. This meant this tasked couldn't be completed in a single or even multiple notebooks, but had to spread across multiple software components.

3.2 Challenges & Solutions

This section will go over the challenges that were encountered over the course of the project. These challenges will be presented in the order, in which we encountered them. I decided on this structure as it introduces these concepts in an understandable manner.

3.2.1 Dealing with arbitrary time-interval-data

The main reason the data was so hard to use, was because of the way time intervals were represented. Often, time-series data is represented by giving measurements over regular intervals. So given a number of time steps $T \in \mathbb{N}$, a number of interval bounds $I \in \mathbb{R}$ and a number of values $V \in \mathbb{R}$ we can define measurement intervals in the following way:

1. For time step $t \in T$ we have the interval bounds $i_t, i_{t+1} \in I$
2. Thusly $|T| = |V| = |I| - 1$ holds
3. With the interval bounds being equidistant if the following holds:
   $|i_t - i_{t+1}| = k; \forall t \in T$

The equidistance defined in 3. is useful, as it can be a base assumption for a number of algorithms and preprocessing techniques. It is also easier to understand.

The intervals in the presented data were not equidistant. This made them hard to understand and evaluate. In figure 1 we can see some of these original measurement intervals visualized. There are measurement intervals, which are only minutes long, while others are more than two days long. Summary statistics like the number of steps taken, the average heart rate or the stress level are thusly reported over different intervals. Dealing with this data in its raw form, would make the data analysis process a lot more complex than it already is,

Let's take another step back and introduce a concept, which is vital for understanding this study. Chemotherapy is administered on a regular basis. The intervals between treatments can however be quite long. Often times, the length between treatments exceeds two weeks. These regular treatment intervals are referred to as "treatment-cycles" or simply "cycles". Since cycles are periods of multiple days, the concept of minutely, irregular measurements does not really apply to them. We decided that the data is better represented in a form of daily measurements. This makes the data a lot less complex to handle and also makes their relationship to cycles more apparent.

To this end, a python module was developed, which took care of "normalizing" the given intervals. This meant either merging intervals together or breaking them apart. This also entailed a recomputation of the relevant measurements, in order to uphold the integrity of the data.

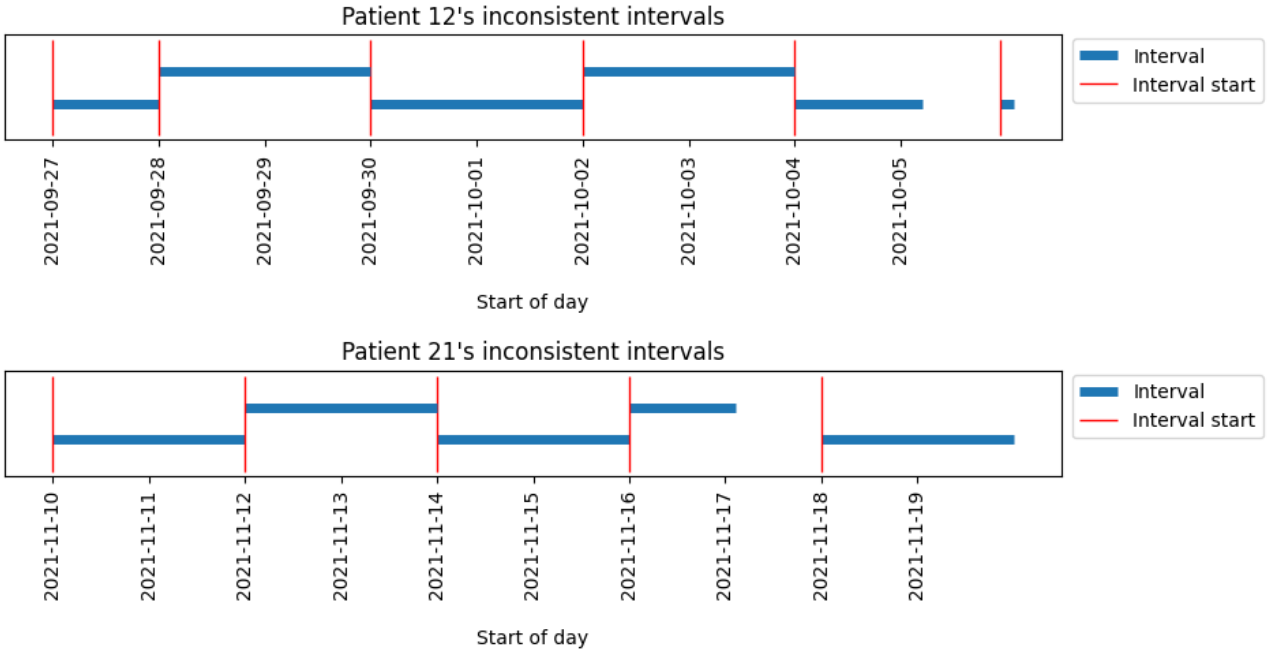

Figure 1: Inconsistent interval bounds visualized

Measurement names	**Aggregation strategy**
User Id, Tracker Id, Group Names	Simple unique aggregator
Duration (s), Steps, Distance (m), Moderate Intensity Duration (s), Vigorous Intensity Duration (s), Floors Climbed, Stress Duration (s), Rest Stress Duration (s), Activity Stress Duration (s), Low Stress Duration (s), Medium Stress Duration (s), High Stress Duration (s)	Fractured sum aggregator
Activity Type, Stress Qualifier	Simple weighted ordinal aggregator
Heart Rate (min bpm)	Minimum aggregator
Heart Rate (max bpm), Stress Level (max)	Maximum aggregator
Heart Rate (avg bpm), Stress Level (avg)	Adjusted mean aggregator

Table 1: Aggregation strategies for all relevant rows

The data aggregation and strategies for each row are specified in table 1. Each of these strategies is a function that takes in a number of intervals and the associated values. Let's say we have the "messy" time steps T' with associated interval bounds I', as well as the equidistant time steps T with bounds I. The function $f(t)$ takes in the messy representation of the time steps and returns the aggregated value for the equidistant time step $t \in T$.

Simple unique aggregator This function simply checks if all the supplied values are the same and returns this single value. It basically only checks the invariant of consistent values across columns for which values that cannot differ. This is the case for the ID variables, since the aggregation happens on a patient by patient basis.

Fractured sum aggregator This aggregator deals with summary statistics that are computed via sums. It deals with merging or dividing these sums across the equidistant intervals.

$$f(t) = \sum_{t' \in T'} I(t, t') \cdot v_{t'}$$

With the indicator function I quantifying the overlap between the equidistant and the "messy" interval.

$$I(t,t') = \begin{cases} 1 & \text{for } i_t \leq i'_{t'} \text{ and } i'_{t'+1} \leq i_{t+1} \\ \frac{i'_{t'} - i_t}{i_{t+1} - i_t} & \text{for } i'_{t'} \in [i_t, i_{t+1}] \text{ and } i_{t+1} < i'_{t'+1} \\ \frac{i_{t+1} - i'_{t'}}{i_{t+1} - i_t} & \text{for } i'_{t'} < i_t \text{ and } i'_{t'+1} \in [i_t, i_{t+1}] \\ 0 & \text{else} \end{cases} \tag{1}$$

Obviously, this method isn't entirely accurate. When an interval spans multiple equidistant time steps, the sum is divided up equally across all the intervals. This makes the assumption that the values were distributed uniformly across the original interval. Obviously, this is not always the case. We still believe that this division is the best course of action, since a final aggregation will look at treatment cycles, rather than individual days. The case that a two-day interval will be divided into two cycles will be rather rare. When it happens regardless, we believe that the overall length of the cycles will give the summed values stability.

Simple weighted ordinal aggregator Some of the values in the data set are ordinal values. These are especially hard to aggregate. Simply appending them or taking the average would ignore the extra information that results from the interval data. Each of the ordinal values is mapped to a normal number representation. We then take the weighted average of all of these values and round to the nearest natural number. Afterwards, we map the retrieved value back to the ordinal scale.

$$f(t) = \lfloor (\sum_{t' \in T'} I(t, t') \cdot v_{t'}) \div (\sum_{t' \in T'} I(t, t')) \rceil$$

With $\lfloor x \rceil$ giving the nearest integer value of $x \in \mathbb{R}$.

Minimum & Maximum aggregator These aggregators look at the minimum or maximum values of an equidistant interval. The function return the maximum value of all "messy" intervals that overlap with the equidistant interval. The function for the maximum aggregator looks like the following:

$$f(t) = \max_{t' \in T'} B(t, t') \cdot v'_{t'}$$

With $B(t, t')$ being a simple indicator function that check if the intervals overlap.

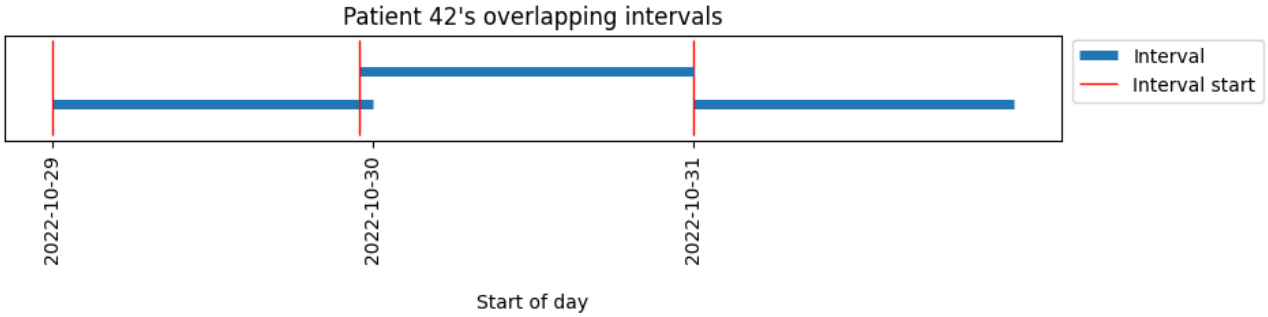

Figure 2: Overlapping intervals visualized

$$B(t,t') = \begin{cases} 1 & \text{if } I(t,t') > 0 \\ 0 & \text{else} \end{cases} \tag{2}$$

A disadvantage of this approach is that it can inflate the maximum and minimum values of the data set. This is because long intervals propagate their maximum or minimum values across many equidistant intervals.

Adjusted mean aggregator The adjusted mean aggregator is very similar to the *weighted ordinal aggregator*. When merging and dividing mean values across equidistant intervals, we want to take interval length into account.

$$f(t) = \frac{\sum_{t' \in T'} I(t,t') \cdot v_{t'}}{\sum_{t' \in T'} I(t,t')}$$

By applying these transformations on the "messy" intervals, I obtained daily measurements, which are much easier to handle.

While applying this process, I even found two intervals in the data set that overlap. Generally, this shouldn't happen, because a patient shouldn't be able to wear their smartwatch twice at the same time. Figure 2 shows that patient 42 has an overlap of 1 hour. This overlap is most likely caused by daylight savings time. We chose to ignore it, since it is the only overlap in a total of 3020 worn days that were recorded.

3.2.2 Visualizing the data

With the data in a daily format, it could be visualized much easier. In order to give a broad overview over the obtained data, an interactive dashboard was developed. This dashboard uses the dash library for python and is accessible over the internet. The dashboard is password protected in order to safeguard the potentially sensitive medical data.

Initially, heatmaps seemed like an intuitive choice to visualize the amount of hours a patient wore their WAD on a given day. One of the initial assumptions was that patients would regularly take off the WAD in order to charge it. This was however disproven once we started operating the dashboard.

Figure 3 shows a screenshot of how the final version of the dashboard looked. On the left there are a number of "patient filters". This data set was gathered and cleaned over the course of the project as well. You can use these multi-select fields to restrict the set of patients displayed in the heatmap on the right. The pie-chart up top shows how many of the patients were filtered out. The dashboard was initially developed to give the medical researchers the ability to define acceptance criteria for the study. These criteria were supposed to restrict the study to patients that wore the WAD an acceptable amount of time. This "acceptability" was still a work in progress, while the dashboard was in development. We developed features that give doctors the ability to define an acceptance criterion for days. This criterion identifies days as valid recordings. This is done by setting a minimum amount of wear time on a daily basis. Furthermore, the researchers can define a minimum number of valid days as well as a

Figure 3: Screenshot of the dashboard in its final state

minimum number of consecutive valid days. The heatmap as well as the pie chart update dynamically based on the values entered for the criteria. Red signifies that a patient was rejected. Green means a patient or a day is accepted. The heatmap shows the accepted days in green, while the other days that contain measurements fade towards black. Completely black days are days, for which no data is present.

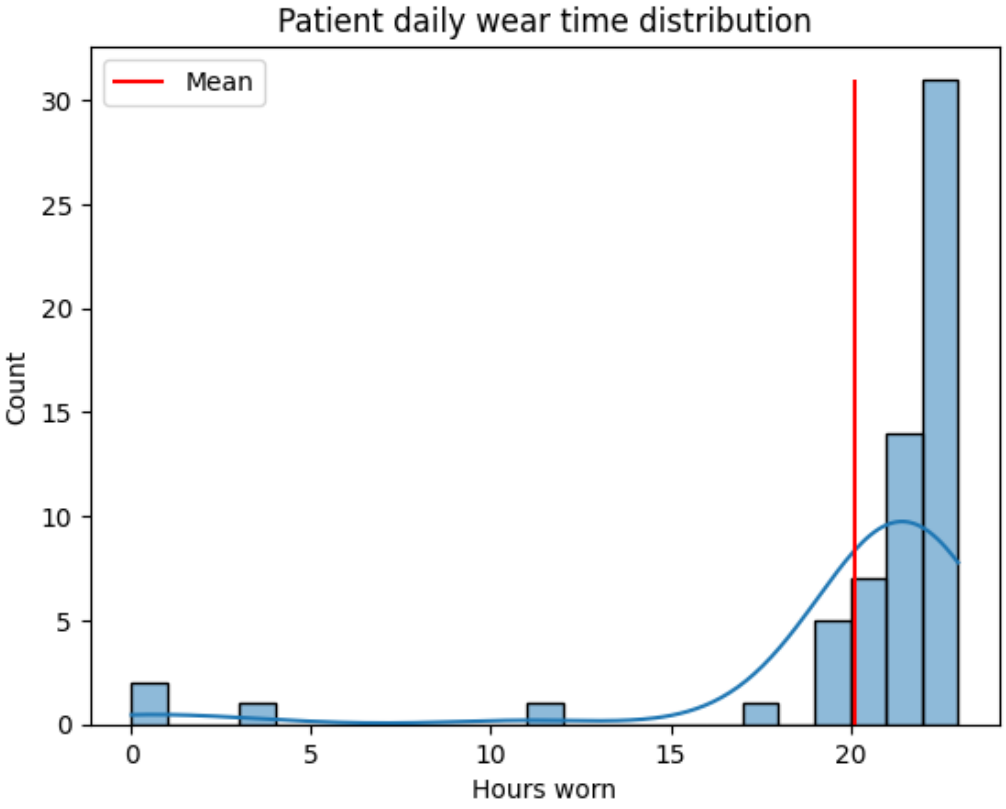

Figure 4: Distribution of wear times on days when patients wore the WAD

We quickly realized that patients either wore the WAD for an entire day or not at all. We can see this distribution visualized in figure 4. This meant that adjusting the daily acceptance criterion did not change much in the overall acceptance rates for patients.

Another important distribution is given by the overall wear time for the patients. It is visualized in figure 5. We can see that many patients wore the WAD under 25 days. There are a few patients that contribute highly to the overall hours worn.

The dashboard clearly showed us that the data is rather sparse. The black spots represent the time across the study period when the watch wasn't worn by the patient. We can see that there are rather large gaps in the data. The reasons for the missing data were already known to the medical researchers. The watches quickly ran out of battery and many patients weren't very proficient with handling them. This meant that most often the WADs were only charged when patients came in for their treatments. The patient did intend to comply with the study, but they simply oversaw charging the watch on a regular basis.

3.2.3 Overcoming data sparsity

As more details about the sparsity of the data were uncovered, we had to come up with a way to deal with this lack of data. This is where the concept of cycles came in handy. We developed the idea of modelling the probability of patients having to postpone or cancel their treatments on a cycle by cycle basis. This means that the smallest unit for modelling would be two treatment-cycles with sufficient data. By looking at the probability of a delay or dropping out of the treatment program, we can identify variables that predict such an event. This cycles-based approach is only possible due to the previous

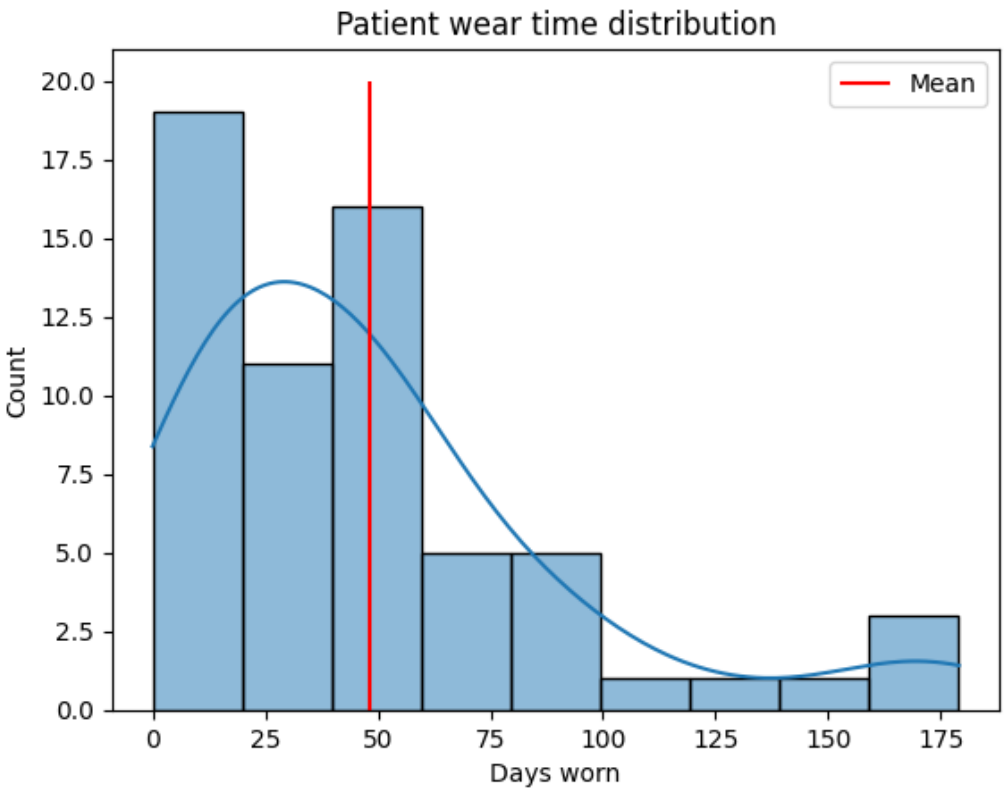

Figure 5: Distribution of overall wear times

restructuring of the data, since intervals used to encompass multiple days. The data now follows a daily format, so we can simply group by cycle and aggregate the variables accordingly.

Column name	Data type
Start date	Date
End date	Date
Type	"Cycle" or "Delay"
ECOG	0, 1, 2, 3, 4, 5
other variables	TBD

Table 2: Treatment cycles data model

We developed a data-format that that will give us the opportunity to gather the treatment cycles and delays for the patients. The data entry task for the cycles data, will be handled by a medicine student, doing their thesis. The data format is outlined in figure 2. This will give us more fine-grained time series data about previously accessible variables like ECOG. We also aim to include other variables in the cycles based model.

4 Insights

In this section, I will go over the research questions and give succinct answers to them based on the previously presented material.

What is the original structure of the data? This question was intended to shed light onto what the original data structure actually meant. The researchers weren't able to work with the unstructured intervals at all. By visualizing the original data, we could see that the intervals varied in length. Many

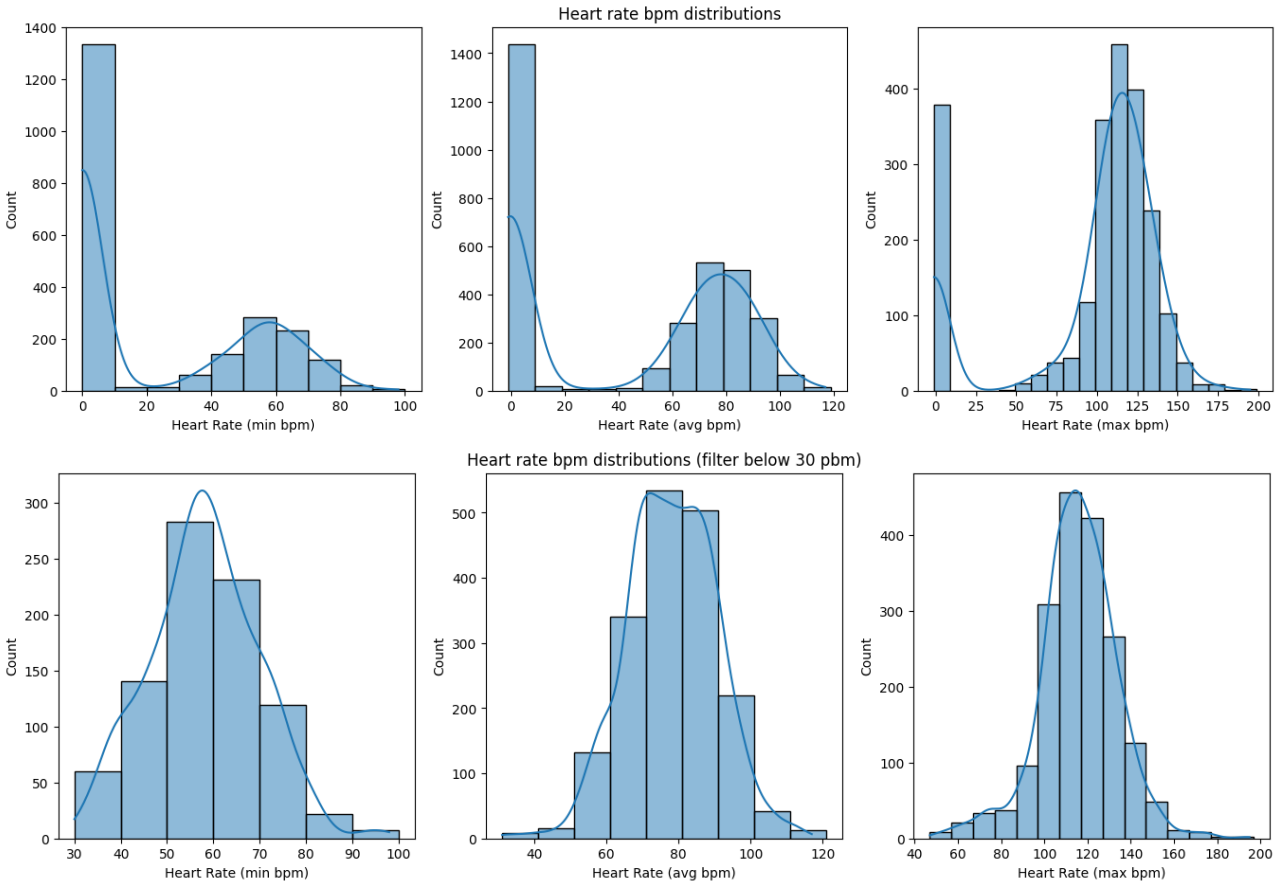

Figure 6: Heart rate distributions visualized before and after cleaning

variables received unintuitive names, like the "Tracker Id" being saved in the "Last Name" field. Understanding the data made it possible to restructure it in a more intuitive way.

In what state should the data be in for it to be of use to the medical researchers? It quickly became apparent, that daily measurements were the most intuitive way to think about the time series measurements. Much of the data contained in the patients logs, kept at AKH tend to specify dates, rather than exact timestamps. Cycles are also structured based on days and not on hours or minutes. This is why we chose a daily structure for the data.

Are the collected measurements actually feasible? When reviewing the data and visualizing its distributions the physicians determined the measurements to be feasible. In figure 6 we can see the heart rate measurements. The data did include some infeasible values, like an average heart-rate of 20 or lower. If only a few of the collected values are infeasible, it is common practice to simply discard them. When filtering out heart rates below 30, the measurements look very reasonable.

What do the medical researchers need in a more general sense to attain the goals of the study? In the beginning, the researchers needed an overview of the data. They needed visualizations and statistics, that gave them a sense of what they are working with. Afterwards, they needed assistance in how to structure the study around the present data.

Did the patients use the WADs on a regular basis, and does this result in a data set of acceptable size? Many of the patients did use the WAD on a regular basis. Often times, patients wear the WAD for an extended period of time, after which there is a considerable gap in the data set. This pattern repeats across the data set.

How can we overcome the sparsity of the data set? We developed a cycles based approach to overcome data sparsity. This means looking at treatment cycles and the probability for a given patient to experience delays or drop out given their previous history.

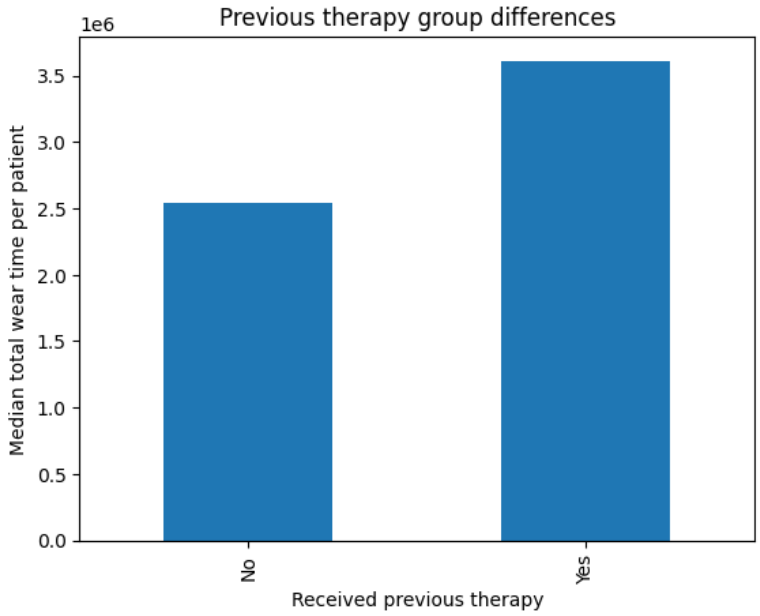

Figure 7: Differing median total wear times per group

Are there any noteworthy patterns in the data? While exploring the data, an interesting pattern was discovered. We observed that patients, already receiving therapy before the project started, generally wore the WAD for a longer period of time. This is visualized in figure 7. This also answers the question *Do any factors like age, type of cancer, type of therapy, gender, etc. impact the wear-time?* We did not find any other significant differences in wear time across the rest of the grouping variables.

5 Conclusion

This project aimed to explain and visualize a complex time series data set concerning regular health related measurements in order to enable further data collection and eventually modelling. By implementing a data complex data transformation module in python, the structure of the data became more intuitive and useful. Many insights into the data and the distribution of the values were attained. An interactive visualization was designed, implemented and deployed in order to give the medical researchers the opportunity to explore the data set for themselves. A data model was developed in order to assist the researchers in the future process of modelling the risk of treatment delays and complications based on regular and fine-grained health data.

This project was a great opportunity to learn more about data science in practice and gain first-hand experience in working with medical researchers. I would be highly interested in pursuing this project further and assist the project team in achieving the stated goal.