Universidad Politécnica de Madrid

Escuela Técnica Superior de Ingenieros Informáticos

MSc. in Data Science

Hepatitis Visualization App

Álvaro García Barragán
Mikel de la Fuente Landa
Maxime Kermagoret

{alvaro.gbarragan,mikel.delafuente,maxime.kermagoret}@alumnos.upm.es

January 7, 2024

Abstract

This application serves as a comprehensive interactive visualization tool designed for medical professionals, specifically doctors, involved in the analysis and monitoring of patients with Hepatitis C. The purpose of the application is to address the unique challenges faced by end-users and analysts when dealing with patient data, offering a solution that enhances their ability to explore, analyze, and understand distinct features within a selected cohort.

Contents

List of Tables

List of Figures

1 Introduction

This report aims to explain the second assignment of the Data Visualization course from the UPM, which consists of the design and implementation of an interactive visualization tool using Shiny[1], a R[2] package that makes it easy to build interactive web applications (apps) straight from R (R is a free software environment for statistical computing and graphics).

Our Shiny application is designed for the visualization of a hepatitis C patient cohort serves several key objectives aimed at enhancing the understanding of this population's health dynamics. The objectives are as follows:

- To enable the selection of specific segments of the patient population based on predetermined characteristics, allowing for a focused analysis on subgroups of interest (henceforth, referred to as "cohort").
- To facilitate the exploration of correlations between specific features, providing insights into how these characteristics may interrelate and impact patient health outcomes.
- To visualize the distribution of particular features, aiding in the identification of common patterns or anomalies that may warrant further investigation.
- To track and display changes in Alanine Aminotransferase (ALT) levels over different time points, illustrating the progression or improvement of liver function as a response to treatment or disease progression.

These objectives are central to the application's purpose, which is to provide healthcare professionals and researchers with a powerful tool for data-driven decision-making and to contribute to the broader knowledge base in the treatment and management of hepatitis C.

1.1 Deploy

The App is published in: https://alvaro8gb.shinyapps.io/HepatitisVisualizationApp/
and the source code can be found in https://github.com/Alvaro8gb/DataVisualizationPW

2 Dataset - Hepatitis C Virus (HCV) for Egyptian patients

Hepatitis is the inflammation of the liver. Inflammation is the swelling of organs that occurs when organs are injured or infected. Inflammation can cause damage to the organs [2]. There are different types of hepatitis, and one type, Hepatitis C, is caused by the Hepatitis C virus (HCV). Hepatitis C can range from a mild illness lasting a few weeks to a severe, lifelong condition.

This study focuses on a dataset [3] comprising 1741 patients infected with the Hepatitis C virus, specifically genotype 4. Patient data was collected at Ain Shams University, Faculty of Medicine, El Demerdash Hospital. These patients underwent treatment with a combined therapy of interferon-Alfa and ribavirin for more than 15 months. The study distinguishes between patients who responded positively to the treatment and those who did not show a clearance of the virus, classified as non-responders.

[1] https://shiny.posit.co/
[2] https://www.r-project.org/

The dataset also provides insights into HCV Liver Fibrosis, encompassing data from patients who received treatment dosages. The "Baseline histological staging" serves as the class label with values {F0, F1, F2, F3, F4}, representing different prognosis levels of Liver Fibrosis:

- No Fibrosis (F0)
- Portal Fibrosis (F1)
- Few Septa (F2)
- Many Septa (F3)
- Cirrhosis (F4)

2.1 Variables Selected for the study

In this section, we outline the features employed in the application. The original dataset comprised 28 features, and from these, we retained only those depicted in Table 1. This feature enhances the dataset utilized by the application, for which we conducted preprocessing to present the features in a more comprehensible manner. For instance, binary features with values 1 and 2 have been mapped with the mapping {1: yes, 2: no}.

Table 1: Selected Features

Feature Name	**Feature Range**	**Feature Type**
Baseline Histological Staging	{F0, F1, F2, F3, F4}	Ordinal
Baseline Histological Grading	$\{n \in \mathbb{Z} \mid 0 \leq n \leq 16\}$	Ordinal
Gender	{Female, Male}	Categorical
Fever	{Yes, No}	Categorical
Nausea/Vomiting	{Yes, No}	Categorical
Headache	{Yes, No}	Categorical
Diarrhea	{Yes, No}	Categorical
Fatigue/Generalized Bone Ache	{Yes, No}	Categorical
Jaundice	{Yes, No}	Categorical
Epigastric Pain	{Yes, No}	Categorical
WBC (White Blood Cell Count)	$[300, 14000]$	Continuous
HGB (Hemoglobin)	$[10, 15]$	Continuous
RBC (Red Blood Cell Count)	$[3 \times 10^6, 5 \times 10^6]$	Continuous
Age	$[32, 61]$	Continuous
BMI (Body Mass Index)	$[22, 35]$	Continuous
Platelet Count	$[9 \times 10^4, 25 \times 10^4]$	Continuous
ALT at 1 month (ALT.1)	$]0, 140[$	Time Series
ALT at 4 months (ALT.4)		Time Series
ALT at 12 months (ALT.12)		Time Series
ALT at 24 months (ALT.24)		Time Series
ALT at 36 months (ALT.36)		Time Series
ALT at 48 months (ALT.48)		Time Series

3 Idioms

In this section, we delineate the idioms implemented in the application along with their respective questions and task abstractions.

3.1 Question 1: How do demographic factors and comorbidities impact the establishment of baseline histological grading and staging?

This question try to see the correlation of different features with baseline histological grading and staging. The selected representation is a Bubble Chart, which is not only used to visualize correlations but also to observe how these correlations are related to the frequency distribution of other features. The different axis, and the size of the bubbles, are to be decided by the user. The features are numerical (BMI, Age, WBC) or categorical (Baseline Histological Staging and Grading).

3.1.1 Data Abstraction

- Attributes: BMI, Age, Baseline Histological Staging, Baseline Histological Grading, WBC
- Ordered attributes: BMI, Age, WBC
- Categorical attributes: Baseline Histological Staging, Baseline Histological Grading

3.1.2 Task Abstraction

Actions:

- Action **Consume:**
  - *Present:* Present to a doctor the impacts of demographic factors and comorbidities on the establishment of baseline histological grading and staging.
  - *Discover:* Examine how variations in age, BMI or WBC can influence baseline histological levels.
- Action **Search:**
  - *Browse:* The user is seeking to understand how white blood cell (WBC) levels are related to demographic factors and comorbidities. The user acknowledges that some links between demographic attributes and baseline histological levels may exist, but doesn't know neither to what extent nor which ones specifically.
- Action **Query:**
  - *Compare:* Compare WBC, BMI and age levels in relation to Baseline Histological Grading and Staging.
  - *Summarize:* Provide an overview of the interactions between WBC, BMI, Age, Baseline Histological Grading and Baseline Histological Staging.

Targets: Identify patterns, trends between some demographic factors and the Baseline Histological Grading and Staging.

Design Choices:

- **Encode:** The visualization employs a Bubble Chart to represent the impact of demographic factors and comorbidities on the establishment of baseline histological grading and staging. The user can choose what attribute he wants on the X-axis as well as on the Y-axis. The size of the bubbles corresponds to the frequency distribution of the selected feature.
- **Manipulate:** Interactive elements, such as dropdown menus, empower users to dynamically filter data based on different demographic factors and comorbidities, like WBC, BMI, Age, Baseline Histological Staging, or Baseline Histological Grading. Users can modify the view, emphasize specific data points, or rearrange the visualization to explore various patterns and outliers.
- **Reduce:** The visualization incorporates reduction techniques, enabling users to filter data according to baseline histological staging or grading, BMI, Age, and WBC levels. This functionality streamlines the view, focusing on pertinent data ranges and potentially aggregating data points for a clearer understanding of trends.

3.1.3 Interaction and Visual Encoding

The user can select what attributes he wants to compare on the Bubble Chart, thanks to dropdown menus that are displayed above the chart. The user can also filter some data thanks to the sliders inputs, in the cohort selection (see section 4).

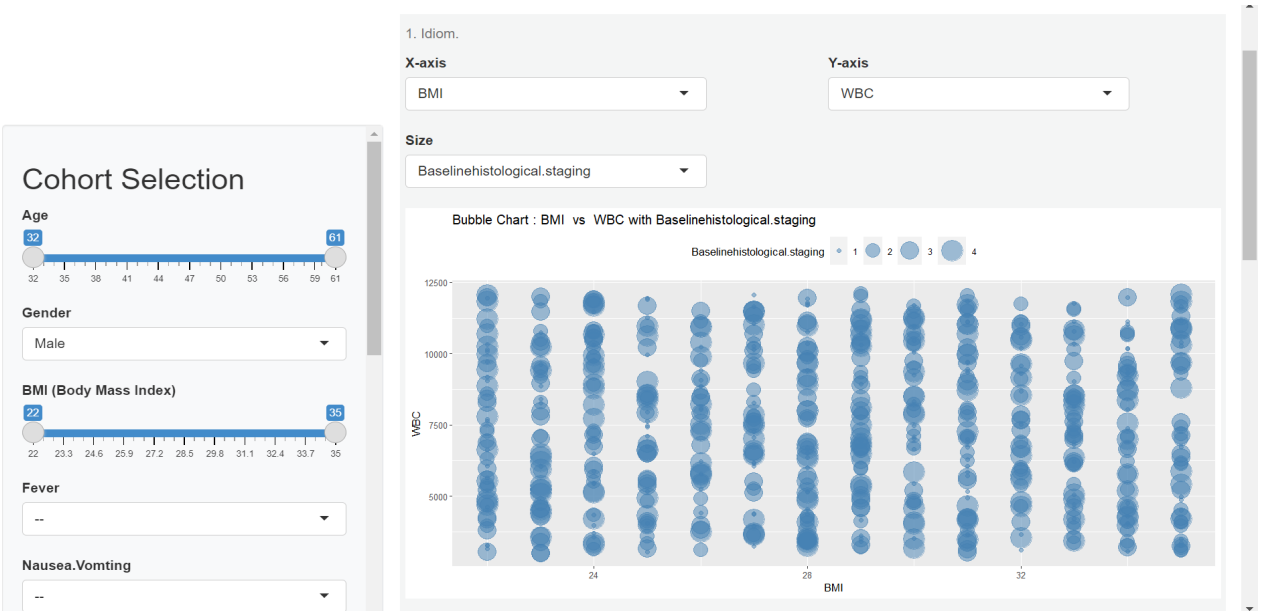

Figure 1: Idiom 1

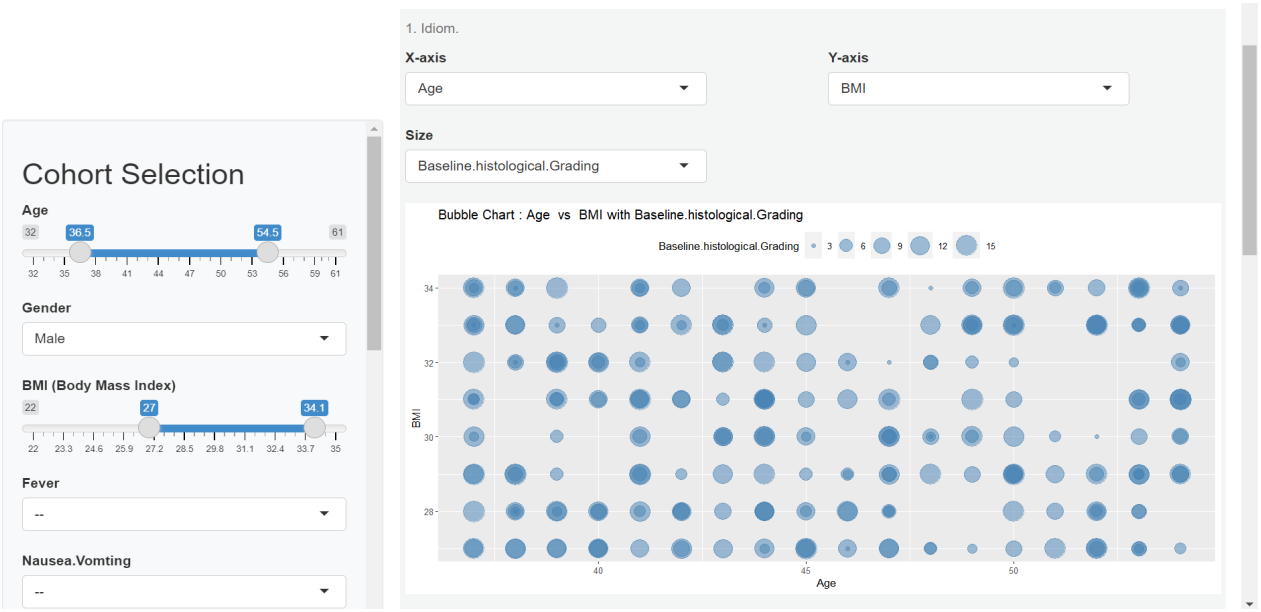

Figure 2: Idiom 1 with sliders

3.2 Question 2: How are demographic factors, comorbidities and the Target variable 'Staging' distributed and how are they distributed grouped by another comorbidity?

This question pertains to the distribution of the different features of the data. The X-axis represents the feature's values, being different for the comorbidities, gender and *Base.line.Staging* (target variable) than for the age and BMI (Body Mass Index), so the Y-axis represents the frequency of cases, being a barplot for those categorical variables and an histogram for the numerical ones.

3.2.1 Data Abstraction

- `Attributes:` Age, Gender, BMI, Fever, Nausea/Vomiting, Headache, Diarrhea, Fatigue generalized bone ache, Jaundice, Epigastric pain
- `Ordered attributes:` Age, BMI
- `Categorical attributes:` Gender (demographic factor), and from the comorbidities: Fever, Nausea/Vomiting, Headache, Diarrhea, Fatigue, Generalized bone ache, Jaundice, Epigastric pain.

3.2.2 Task Abstraction

Actions:

- Action **Consume:**
  - *Present:* Display the distribution of demographic factors, comorbidities and the target variable, considering also another feature if desired.
  - *Discover:* Identify patterns and trends in the distribution of features, as well as patterns between two features.

- Action **Search:**
  - *Browse:* Users know they can navigate through the different data cases, but don't know exactly what they are looking for.
- Action **Query:**
  - *Identify:* Identify the characteristics of the distribution of a feature.
  - *Compare:* Compare the distribution of different features.

Targets:
Find trends and patterns of demographic factors and comorbidities.

Design choices:

- **Encode:** Barplots for categorical variables visually express the frequency or count of data by representing each category as a separate bar along the x-axis; for the ordered attributes the histograms express the distribution of numerical data by representing the frequency or count of data points within the range of predefined intervals. The grouped variable represents by colour the frequency in the barplot or in the histogram.
- **Manipulate:** A main dropdown and the group feature menu allow for dynamic visualization of the different features. This manipulation enables users to change the view, highlight specific data features and examine different groupings, of just a specific feature or a pair of them.

3.2.3 Interaction and Visual Encoding

The user selects the feature they want to analyze from the options displayed in the dropdown menu of the interface. The second dropdown menu presents the variable by which the data will be grouped. If the same variable is chosen, a simple distribution is displayed.

The plot changes based on the type of feature:

- **Categorical**: A barplot is used for this case, Figure 3 shows an example of the visualization for a simple distribution of a feature.
- **Ordinal**: An histogram is used for this case, Figure 4 shows a grouped distribution of two different features. shows an example of the visualization within another categorical variable.
- **Numerical**: An stacked histogram is used for this case, Figure 5 shows an example of the visualization within another categorical variable.

This visualization aims to represent the distribution in the most effective manner. Therefore, we avoid using a stacked plot when the number of distinct instances of the grouped feature is greater than 2. This is done to ensure that groups can be compared in a meaningful way.

Figure 3: Distribution of Headache feature

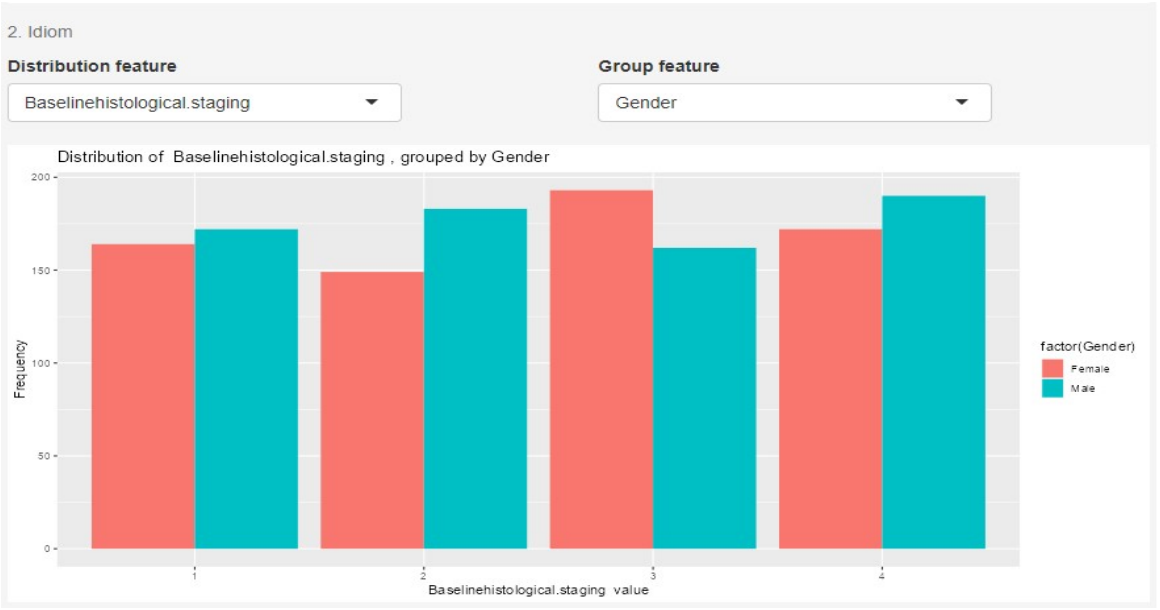

Figure 4: Distribution of the target feature, grouped by Gender feature

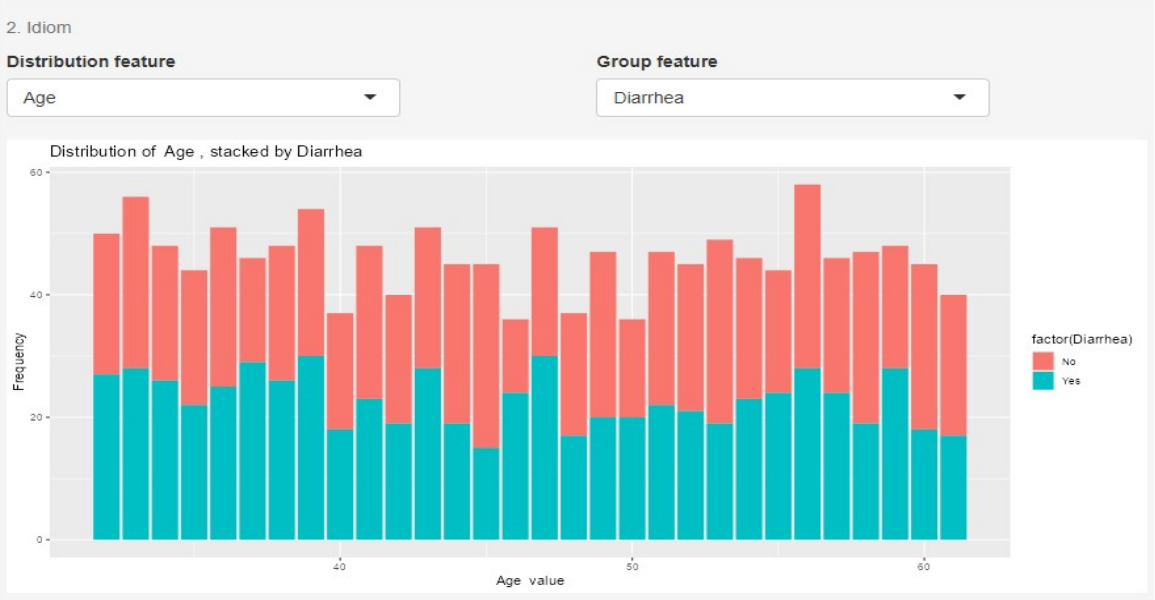

Figure 5: Distribution of Age feature, stacked by Diarrhea feature

3.3 Question 3: What is the progression of ALT levels at 1, 4, 12, 24, 36, and 48 weeks in the patients based on their baseline histological staging levels?

Given that this question pertains to temporal data, which is essentially a time series, we opt to represent it through visualization in the form of a line chart. The y-axis denotes the quantitative value of ALT (alanine transaminase), while the X-axis represents the respective ordered time intervals during which the measurements are taken.

3.3.1 Data Abstraction

- `Attributes:` ALT levels in different times and Baseline histological staging.
- `Ordered attributes:` ALT levels in different times.
- `Categorical attributes:` Baseline histological staging.

3.3.2 Task Abstraction

Actions:

- Action **Consume:**
  - *Present:* Present to a doctor the temporal changes in ALT levels
  - *Discover:* Examine whether patients with varying baseline histological levels exhibit disparities in the temporal patterns of ALT levels.
- Action **Search:**
  - *Browse:* The user is familiar with the significance of ALT levels but lacks information about the patients.
- Action **Query:**

- *Compare:* Compare the levels of ALT between patients.
- *Summarize:* Overview the media, the standard deviation range, and the minimum and maximum recommended levels of ALT.

Targets: Identify trends, patterns and outliers.
Design choices:

- **Encode:** The graph uses line marks to track individual data points over time, with color fills indicating variance, and a dashed line to represent average trends. Spatial arrangement allows for comparison across different stages of ALT levels, directly expressing values through vertical positioning on the graph. Additionally, the graph encodes the recommended minimum and maximum values for ALT, obtained from the literature [1].
- **Manipulate:** Interactive elements such as a dropdown menu allow for dynamic filtering of data based on baseline histological staging. This manipulation enables users to change the view, highlight specific data paths, or reorder the visualization to examine different patterns or outliers.
- **Reduce:** The visualization implements reduction techniques by allowing users to filter data according to histological stages. This can simplify the view to focus on relevant data ranges, potentially aggregating data points to sort the graph and facilitate a clearer understanding of trends.

3.3.3 Interaction and Visual Encoding

The user is prompted to select a baseline histological staging between 1 to 4. The line chart then visualizes the temporal pattern distribution of patients within the chosen cohort, complete with average (red dashed line), standard deviation (light green shape), an min (yellow dashed line) and max (green dashed line) values as depicted in Figure 7. To avoid overloading the graph with information, the minimum and maximum values could be hidden (Figure 6).

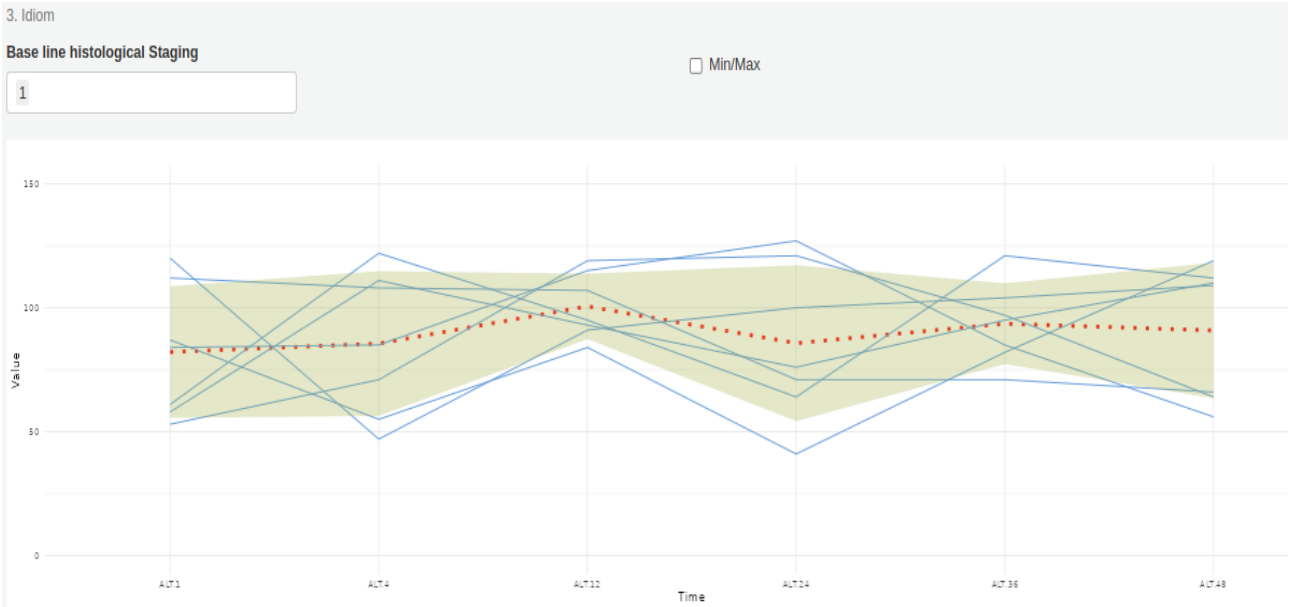

Figure 6: Idiom 3

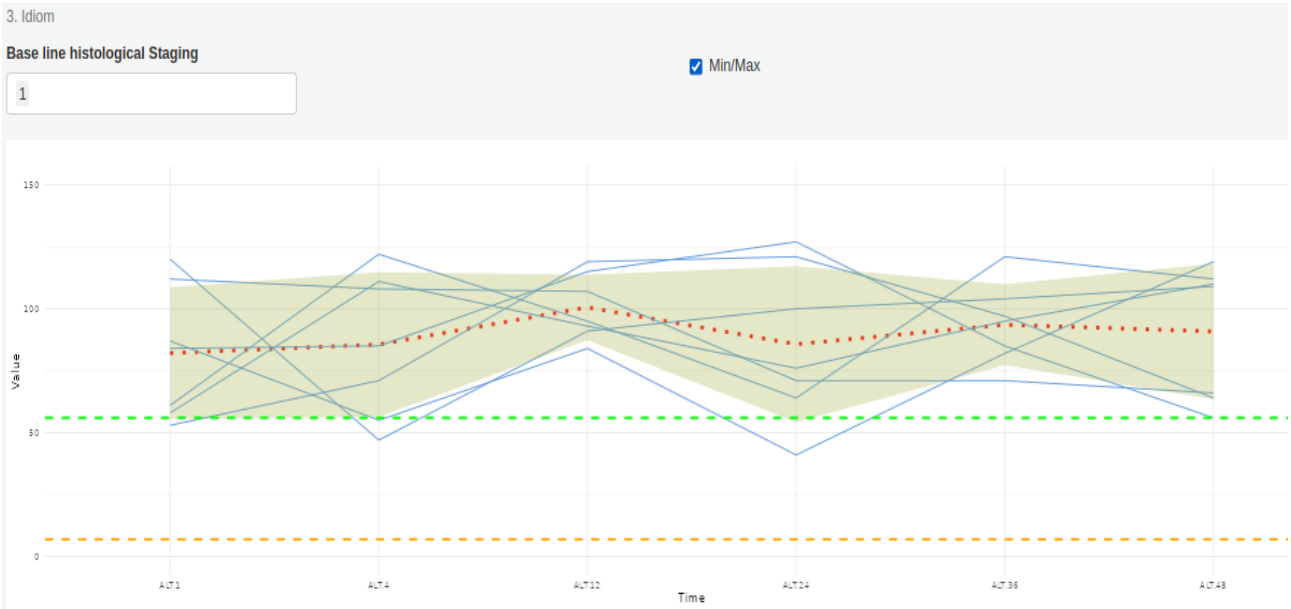

Figure 7: Idiom 3 with min-max

4 Cohort Selection Interface

The interface employs synchronized multi-view filtering to facilitate a more streamlined data processing experience. The 'Cohort Selection' panel serves as a central hub for refining patient data, ensuring that each view consistently reflects the selected parameters. Users can adjust the following filters to narrow down the patient cohort:

- **Age Range:** A slider allows the selection of a patient age range, enhancing focus on a specific age group.
- **Gender:** Options to select male, female, or both genders are available for inclusive data analysis.
- **BMI (Body Mass Index):** Another slider to select a BMI range, targeting patients within a particular weight category.
- **Symptoms Filtering:** A collection of symptoms, including fever, nausea, headache, etc., can be selected based on whether an patient has, does not have, or do not filter.
- **Blood Counts:** Sliders for white blood cell count, red blood cell count, hemoglobin levels, and platelet count provide a detailed hematological filtering mechanism.

This granular approach to data selection enables researchers to isolate and examine subsets of the patient population based on specific criteria, thus making the analysis more manageable and focused as depicted in Figure 8.

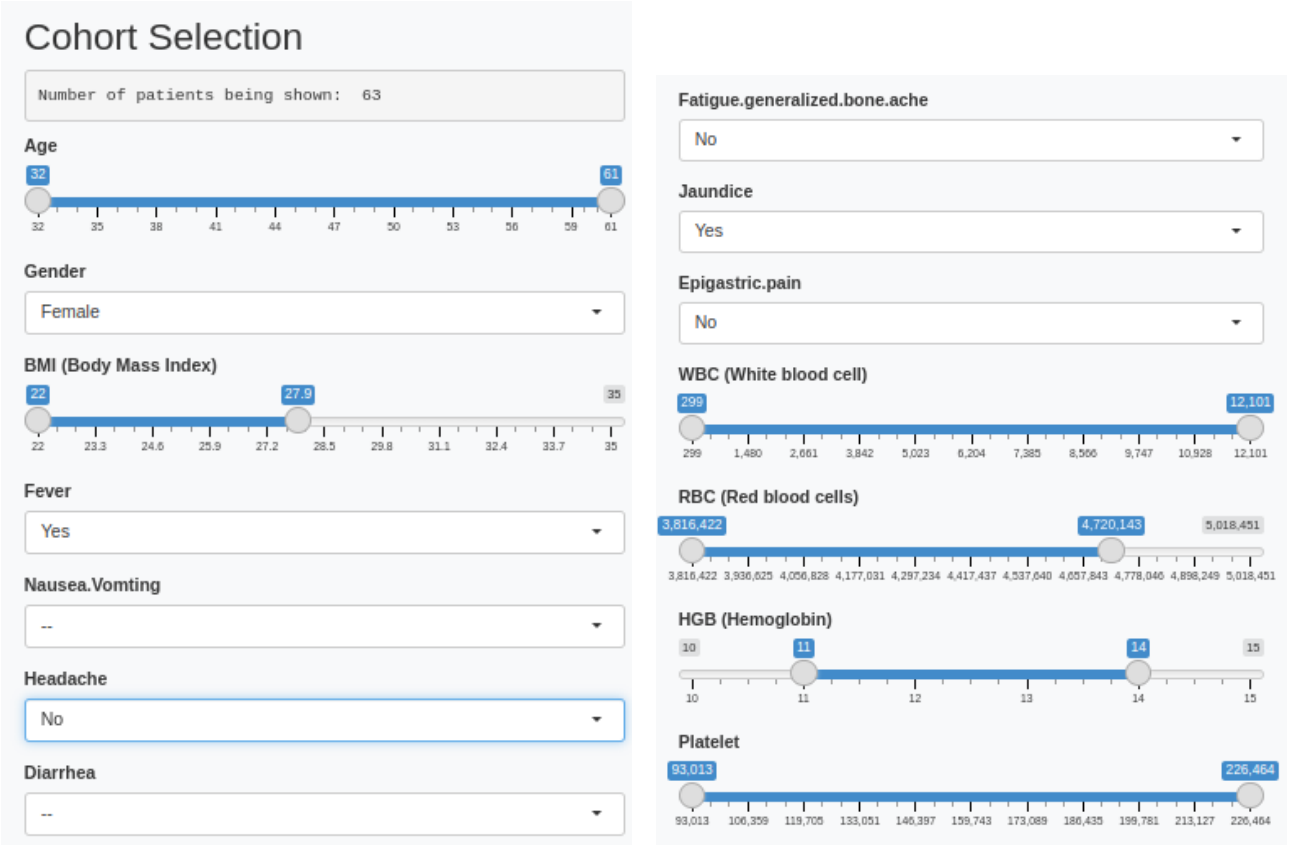

Figure 8: Cohort selection interface.

5 Instructions on how to use the App

Follow these instructions to get your Shiny app up and running:

1. **Make Sure R and Shiny are Installed:**
   - Ensure that you have R installed on your machine.
   - Install the Shiny package if you haven't already by running the following command in the R console:

```
install.packages("shiny")
install.packages("dplyr")
install.packages("RColorBrewer")
install.packages("ggplot2")
install.packages("stringr")
```

2. **Set Permissions for the Launch Script (launch.sh):**
   - Make sure the launch script is executable. You can set the execute permission with the following command:

```
chmod +x launch.sh
```

3. **Run the Launch Script:**
   - Execute the launch script from the terminal:

```
./launch.sh
```

4. **Alternatively, Run Directly from R Console:**
   - Run the following commands:

     ```
     Rscript app.R
     ```

     This will directly run your Shiny app without using the launch script.
5. **Access the App in a Web Browser:**
   - After running the app, it should provide an output with a URL. Usually, it's something like `http://127.0.0.1:port/`.
   - Open this URL in your web browser to view and interact with your Shiny app.

Make sure to resolve any dependency issues if they arise. If there are any specific error messages or issues you encounter, feel free to ask for further assistance.

6 Conclusion

In conclusion, our Shiny-based interactive visualization tool for Hepatitis C patient data offers a comprehensive solution for medical professionals, providing enhanced capabilities to explore, analyze, and understand distinct features within a patient cohort. Developed as part of a Data Visualization course, the tool systematically addresses the assignment requirements, utilizing a dataset of 1741 Egyptian patients treated for Hepatitis C.

The application incorporates diverse visualizations, including bubble charts, bar plots, histograms and line charts, to fulfill specific objectives such as exploring correlations, identifying patterns, and tracking temporal changes in ALT levels. The interactivity of the tool allows users, particularly doctors, to dynamically filter and aggregate data, facilitating personalized analyses based on demographic factors, comorbidities, and other criteria.

The cohort selection interface serves as a key feature, enabling streamlined data processing by allowing users to isolate subsets of the patient population based on age, gender, BMI, symptoms, and blood counts. This granular approach enhances the utility of the tool for healthcare professionals and researchers, contributing to data-driven decision-making and advancing knowledge in Hepatitis C treatment and management.

To finish with, our Shiny application surpasses minimum requirements by incorporating additional design options, ensuring a user-friendly experience. Its potential impact lies in presenting complex patient data intuitively, offering valuable insights for informed decision-making in the healthcare domain.

7 References

[1] *Alanine Amino Transferase* at: https://www.ncbi.nlm.nih.gov/books/NBK559278/ (Accessed: 26/12/2023).

[2] *Hepatitis* https://www.ncbi.nlm.nih.gov/books/NBK554549/ (Accessed: 26/12/2023)

[3] *Hepatitis C Virus (HCV) for Egyptian patients*

https://archive.ics.uci.edu/dataset/503/hepatitis+c+virus+hcv+for+egyptian+patients (Accessed: 26/12/2023)