MASTER THESIS

Visual Analysis for Narcolepsy

This thesis is submitted in partial fulfillment of the requirements for the degree of Master of Science

in

Computer Science

by

Priyanka Bhaskar
Student Number: 4772989

under the supervision of

Prof. Dr. Anna Vilanova
Dr. Michel Westenberg

Abstract

Visual Analysis for Narcolepsy

by

Priyanka Bhaskar

Narcolepsy is a chronic neurological condition that results from the dysregulation of the sleep-wake cycle occurring in an early stage, specifically in adolescence. Patients with narcolepsy experience excessive daytime sleepiness, cataplexy, hypnagogic hallucinations, sleep paralysis and disturbed nocturnal sleep and these symptoms together form the narcolepsy symptom pentad. However, the symptoms related to narcolepsy are not limited to the pentad and cover a broad range of other symptoms, some of which are not directly related to sleep, like, increase in weight, binge eating, anxiety, agitation. Therefore, researchers from sleep medicine center, Kempenhaeghe wanted to understand how a selected set of 20 symptoms are related to narcolepsy.

In this thesis we present a visual analytics framework to help the researchers understand these symptoms in relation to narcolepsy and identify patterns among them. Using relevant attributes interesting population subsets are formed and the results are compared. The thesis comprises of three main tasks, the first being visualizing the distribution of individual attributes of patients in which a symptom is present and not present and for severity level. Then, pairs of symptoms are visualized based on their agreement and association to identify groups of related symptoms and trends present. Lastly, multiple symptoms are visualized to identify patterns amongst them.

The visual analytics framework was implemented and evaluated through a user study. The study showed that the visualization developed was useful in gaining new insights to form interesting hypothesis along with possible extensions for future development. This is the first step towards an extensive visual analytics framework in the study of narcolepsy symptom spectrum.

Thesis Committee:

Prof. Dr. Anna Vilanova - TU Delft - TU Eindhoven
Dr. Willem-Paul Brinkman - TU Delft
Dr. Thomas Höllt - TU Delft

Acknowledgement

The thesis project has been an extremely rewarding personal journey and it has given me the opportunity to expand my application knowledge. Working on such a complex task wouldn't have been possible without the support and encouragement of various stakeholders who have helped shape my thoughts, encouraged me from time to time, given me useful insights, and most importantly given me the freedom to stumble and learn while working on the project. I would now like to take this opportunity to individually thank the following people for their valuable contribution and whose support has been crucial while working on the project.

Prof. Dr. Anna Vilanova, intially when I approached her she took her valuable time out to understand my requirements and explained to me the limited projects that she had to offer but was kind enough to introduce me to professors at Tu Eindhoven who had projects to my liking. She has been extremely gracious in sharing her vast experience, giving me direction, providing technical inputs when I was in doubt and patiently providing feedback on the various design iterations and reports. I would also like to thank her for the personal time she set aside for our regular weekly meetings and her uncompromising approach to bring out the best in the project.

I initially started the project under the guidance of Prof. Dr Michel Westenberg who guided me in setting up the objectives and understanding the target users expectations. During my interactions with Prof. Michel he gave me the freedom to explore the various choices before settling on the best course of action. He was calm with a friendly demeanor and I thank him for the valuable time spent with me during our weekly interactions.

Dr. Sebastiaan Overeem is one of the core member from the sleep medicine center, Kempenhagen and his deep understanding of the subject and inputs have been invaluable in shaping the project. He is very insightful, quick in his response, clear in his expectations, available for discussions on short notice and has been constructive in his feedback. I thank him for his inputs and constant encouragement while working on the project.

Dr. Sigrid Pillen is one of the core members from the sleep medicine center, Kempenhagen. Her various inputs and suggestions have helped me clarify my doubts and gain a better perspective of the subject. I thank her for participating in our discussions and providing her valuable feedback.

Drs. Laury, who is pursuing her doctorate in psychology at the sleep medicine centre, Kempenhagen, has been very helpful in providing me with the necessary data with regular updates and clarifying all my doubts. I would like to thank her for participating in our discussions and providing me with her valuable feedback.

I would like to extand my thanks to the thesis committee, Dr. Willem-Paul Brinkman and Dr. Thomas Höllt for having taken time off their busy schedule to be a part of the committee.

Marjo van Koppen, core member of TU Delft Introduction program. She excudes confidence, is very jovial, and has constantly followed up on my project progress. She has gone out of the way to provide me support for which I remain indebted. I would also like to thank my team members from the Introduction program, for giving me the best

experience one could have asked for, helping me break the isolation that one encounters and being my stress busters.

Lastly, this list wouldn’t be complete without thanking my parents for their constant support, motivation and encouragement. They have played an immense role in this project and in this entire master’s journey and without their support this wouldn’t have been possible.

Contents

List of Figures ix

List of Tables xi

1 Introduction 1
1.1 Motivation . . . 1
1.2 Problem Statement . . . 3
1.3 Contribution . . . 4
1.4 Report Organisation . . . 5

2 Related Work 6

3 Domain Analysis 11

4 Task and Data Abstraction 14
4.1 Data Abstraction . . . 14
4.2 Task Abstraction . . . 14

5 Visual encoding and Implementation 16
5.1 Tool used for Implementation . . . 16
5.2 Overview . . . 16
5.3 Filtering - Subset Selection . . . 17
5.3.1 Filtering of records . . . 17
5.3.2 Subset selection . . . 18
5.4 Visualization of individual attributes . . . 22
5.4.1 Distribution of symptoms present and not present . . . 22
5.4.2 Distribution of symptom severity and identification of trends . . . 23
5.5 Visualization of pairs of attributes . . . 24
5.5.1 Co-occurrence of symptoms based on reporting . . . 24
5.5.2 Relation of symptoms based on severity . . . 25
5.5.3 Identify groups and trends . . . 27
5.6 Visualization of multiple attributes . . . 31
5.7 Comparison of population subsets . . . 34
5.7.1 Comparison of individual symptoms . . . 36
5.7.2 Comparison of pairs of symptoms . . . 38
5.7.3 Comparison of multiple symptoms . . . 42
5.8 Dashboard . . . 43
5.9 Implementation . . . 45

6 Evaluation **47**
6.1 Method and participants 47
6.2 Results 48
6.2.1 Selection mechanism 48
6.2.2 Visualizing individual symptoms 48
6.2.3 Visualizing pairs of symptoms 49
6.2.4 Visualizing multiple symptoms 49
6.2.5 Comparison 50
6.2.6 Overall visualization 50

7 Conclusion **51**
7.1 Summary 51
7.2 Future work 52

A Dashboard **53**

B User Study Questionnaire **55**

C Complete User Study Responses **61**

Bibliography **64**

List of Figures

1.1 Structure of the Mobile application 2
1.2 Mobile application interfaces 3
1.3 Tamara Munzner four nested levels of visualization. (Source: Visualization Analysis & Design, Tamara Munzner) 4

2.1 LifeLines2 - all patients records are aligned based on the 1st reported pneumonia and influenza on a relative scale. (Source: Aligning Temporal Data by Sentinel Events: Discovering Patterns in Electronic Health Records) 7
2.2 Interface of TripVista. (Source: TripVista: Triple Perspective Visual Trajectory Analytics and Its Application on Microscopic Traffic Data at a Road Intersection) 9

3.1 Snapshot of the dataset 12

5.1 Overview of tasks 17
5.2 Filtering of records when timestamp entered is 50 hours 18
5.3 Distribution of number of times reported by patients using the mobile application 19
5.4 Number of times reported among those reporting more than twice 20
5.5 Age 21
5.6 Ullanlinna score 22
5.7 Stacked bar plot depicting the distribution of symptoms in terms of those reporting and not reporting a symptom 23
5.8 Individual bar plots depicting the distribution of symptoms in terms of severity . . 23
5.9 Balloon plot depicting the number of patients reporting different combinations of reporting and not reporting 24
5.10 Balloon plot depicting the number of patients reporting different combinations of symptom severity 25
5.11 Toy dataset to illustrate kendall tau b for symptoms anxiety and agitation 27
5.12 Illustration of kendall tau b for the toy dataset 27
5.13 Heatmaps for agreement and association between pairs of symptoms 28
5.14 Heatmaps ordered based on value 29
5.15 Heatmaps ordered based on cluster 29
5.16 Matrix representation of cluster formation on a toy dataset 30
5.17 Dendrogram formed for the toy dataset 31
5.18 Parallel coordinate plot 32
5.19 Parallel coordinate plot - left plot for interaction and the right plot to reflect the result of interaction 33
5.20 Parallel coordinate plots when hovered on the plot on the left 33
5.21 Parallel coordinate plots when brushed on the plot on the left 34
5.22 Parallel coordinate plots when brushed and hovered on the plot on the left 34
5.23 Three methods of faceting. (Source: Pep up Your Time Machine – Recommendations for the Design of Information Visualizations of Time-Dependent Data) 35
5.24 Stacked bar plots ordered based on their individual ordering 36

5.25 Stacked bar plots ordered based on the left hand side plot's order 36
5.26 Stacked bar plots ordered based on the right hand side plot's order 37
5.27 Individual bar plots depicting the distribution of symptoms in terms of severity with absolute scale 37
5.28 Individual bar plots depicting the distribution of symptoms in terms of severity with relative scale 38
5.29 Individual bar plots depicting the distribution of symptoms with symptoms highlighted corresponding to the symptoms selected in the stacked bar plot. 38
5.30 Heatmaps reordered using option - Individual 39
5.31 Heatmaps reordered using Agreement-left 39
5.32 Heatmaps reordered using Association-left 40
5.33 Heatmaps reordered using Agreement-right 40
5.34 Heatmaps reordered using Association-right 41
5.35 Result when hovered over the heatmaps 41
5.36 Balloon plots pertaining to two different subsets 42
5.37 Parallel coordinate plots for the two different subsets selected 42
5.38 Filtering and Subset selection 43
5.39 Individual and multiple symptoms 44
5.40 Pairs of symptoms 45

B.1 User study questionnaire - part 1 of 11 55
B.2 User study questionnaire - part 2 of 11 55
B.3 User study questionnaire - part 3 of 11 56
B.4 User study questionnaire - part 4 of 11 56
B.5 User study questionnaire - part 5 of 11 57
B.6 User study questionnaire - part 6 of 11 57
B.7 User study questionnaire - part 7 of 11 58
B.8 User study questionnaire - part 8 of 11 58
B.9 User study questionnaire - part 9 of 11 59
B.10 User study questionnaire - part 10 of 11 59
B.11 User study questionnaire - part 11 of 11 60

C.1 Key for user study responses 61
C.2 User Study Responses - Selection Mechanism 61
C.3 User Study Responses - Visualizing individual symptoms 62
C.4 User Study Responses - Visualizing pairs of symptoms 62
C.5 User Study Responses - Visualizing multiple symptoms 62
C.6 User Study Responses - Comparison 63
C.7 User Study Responses - Overall visualization 63

List of Tables

3.1 Table depicting the list of attributes and symptoms collected via the questionnaire interface 12

3.2 Table depicts the list of 20 symptoms from which patients can select and rank the symptoms experienced. 12

4.1 Table briefly depicts the attributes falling under each data type 14

6.1 Color scheme used to represent the results obtained 48

6.2 User study questions and responses for selection mechanism 48

6.3 User study questions and responses for visualizing individual symptoms 48

6.4 User study questions and responses for visualizing pairs of symptoms 49

6.5 User study questions and responses for visualizing multiple symptoms. 49

6.6 User study questions and responses for comparing two population subsets. 50

Acronyms

KNAVE-II Knowledge-based Navigation of Abstractions for Visualization and Exploration-II. 8

PRIMA Patient Record Intelligent Monitoring and Analysis. 6

Chapter 1

Introduction

1.1 Motivation

Narcolepsy is a chronic sleep disorder with a typical onset in adolescence, characterized by excessive daytime sleepiness and cataplexy [1][2]. Apart from experiencing daytime sleepiness and cataplexy, patients with narcolepsy can also experience disturbed night time sleep, sleep paralysis and hallucinations (hypnagogic or hypnopompic) [1][2]. The five symptoms together form what is known as a narcolepsy symptom pentad. However, the symptoms related to narcolepsy are not restricted to the symptom pentad and can range from sleep related to non-sleep related like binge eating, agitation, increase in weight to name a few.

Narcolepsy is said to occur 1 in 2000 people worldwide, and delay in diagnosis is common [3]. Reasons for the delay in diagnoses include, physicians lack of familiarity with narcolepsy, lack of access to laboratory-based sleep testing for the patients, and a broad variety of both medical and psychiatric comorbidities associated with narcolepsy, such as obesity, other sleep disorder, depression and anxiety [2][4][5]. In general symptoms develop over time and initial diagnostic tests may be falsely negative despite severe and specific clinical symptoms and signs [6]. Patients with narcolepsy face challenges such as social stigma, difficulty in obtaining an education and maintaining a job, reduced quality of life and socio-economic consequences [1][6]. A few ways that will aid in the quicker identification of the onset of narcolepsy include improving awareness about the diagnosis and tailored therapies, initiating early medication when cataplexy is present which is a key diagnostic marker and considering multiple concurrent medications [3][6].

Symptoms can vary amongst patients and, day to day variations are also possible. Some of the factors that influence this variation are, medication and medication tolerance, circumstances of daily living and variation in coping strategies. Currently, the diagnostic tests focus on 2 core symptoms namely cataplexy and excessive day time sleepiness, and it is found that questionnaires are unable to cover the whole symptom spectrum and its long term day to day variability. This lead the researchers at sleep medicine centre, Kempenhaeghe, to build a mobile application that covers all the symptoms and allows the users to log them on a timely basis as and when they experience it.

The mobile application is available in both Android and IOS and lets users, log their severity towards a comprehensive list of twenty symptoms for narcolepsy. This not only lets the users keep track of their symptoms but also serves as a source of dataset that can be used for analytics purpose. The results can help doctors in gaining more insight which aids in identifying the onset of narcolepsy.

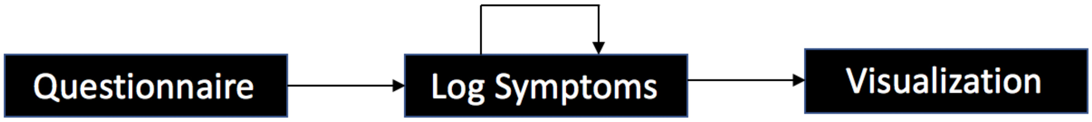
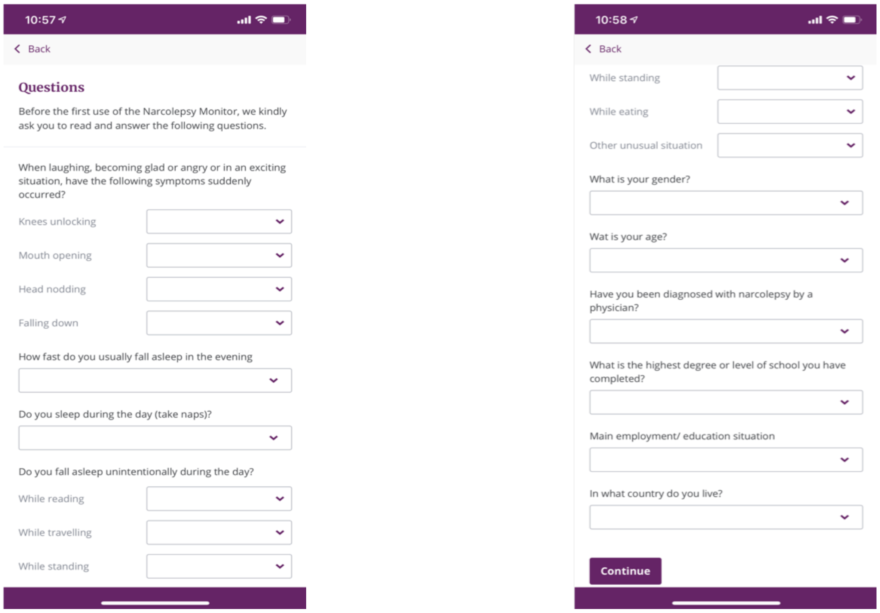


Figure 1.1: Structure of the Mobile application

The application is structured as shown in Figure 1.1, where when a user logs in for the first time they will have to fill in a questionnaire which helps in calculating the Ullanlinna Score. A Ullanlinna score of 14 and above is suggestive of narcolepsy with cataplexy [7]. After filling the questionnaire, the users get to select symptoms they are experiencing and report severity. The severity of symptoms is asked instead of frequency as it is believed that knowing the impact a symptom has on patient gives a deeper understanding of narcolepsy. The app also provides its users with a basic visualization to review their complete reporting. Figure 1.2a and 1.2b depicts the questionnaire interface, followed by Figure 1.2c which depicts the interface where users can log their symptoms and the severity level. Figure 1.2d depicts the visualization provided in the mobile application for the users to view their reporting patterns.



Questions

Before the first use of the Narcolepsy Monitor, we kindly ask you to read and answer the following questions.

When laughing, becoming glad or angry or in an exciting situation, have the following symptoms suddenly occurred?

Knees unlocking

Mouth opening

Head nodding

Falling down

How fast do you usually fall asleep in the evening

Do you sleep during the day (take naps)?

Do you fall asleep unintentionally during the day?

While reading

While travelling

While standing

(a) Questionnaire Interface - 1

While standing

While eating

Other unusual situation

What is your gender?

Wat is your age?

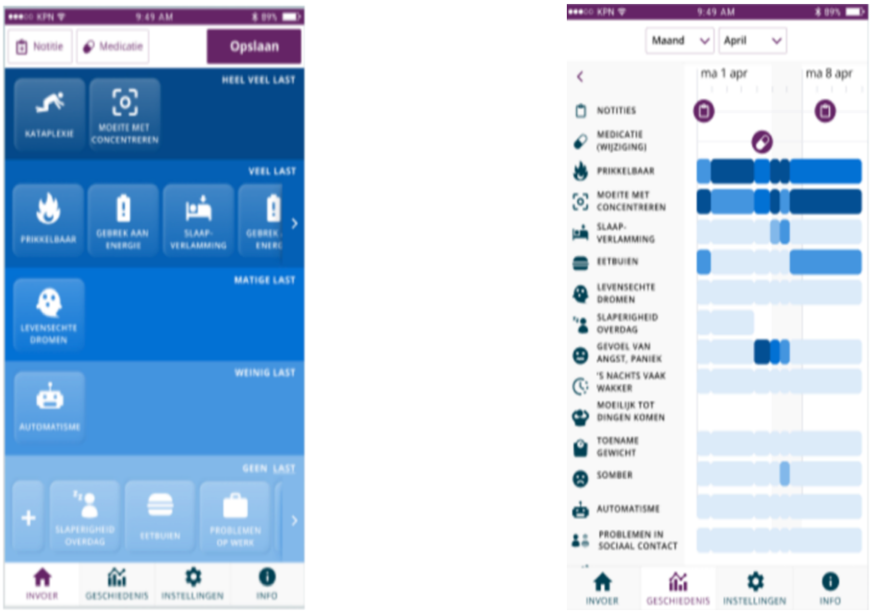
Have you been diagnosed with narcolepsy by a physician?

What is the highest degree or level of school you have completed?

Main employment/ education situation

In what country do you live?

(b) Questionnaire Interface - 2



(c) Interface for patients to log symptoms and its severity

(d) Visualization provided to the patients to see their reporting patterns

Figure 1.2: Mobile application interfaces

1.2 Problem Statement

From the data collected through the mobile application which comprises of personal, demographic and clinical symptoms with severity levels, there are two main focus areas. One is to understand the symptom spectrum better and the other is to understand individual patients reporting in detail. Time aspect of the dataset plays a crucial role when the focus is on individual patients, when the focus is on the overall symptom spectrum the time component can be ignored as here we are primarily focusing on the relations between the symptoms and not with respect to a time stamp. After having detailed discussions with the researches from the sleep medicine center, Kempenhaeghe it was decided that this project will focus on understanding the symptom spectrum. This knowledge will help when reviewing the records of individual patients to understand the reporting pattern better and to give a personalized treatment.

The goal of this project is to help the researches get an understanding of the dataset population using patients personal details and reporting patterns. It enables to identify patterns present among the symptoms and the relations to narcolepsy to facilitate the formulation of hypothesis.

In order to explore this growing dataset, as existing and new patients keep reporting the symptoms experienced and the associated severity, there is a need to effectively present this data. This can be achieved with the help of visualization to depict information which can be grasped quickly when compared to manual data search or use of tables. Since, there is no clear model defined using appropriate visualization helps in easily identifying relationships and patterns if present and also helps compare the visualizations for different subsets of the dataset population effectively. Visualization also supports analysis at various levels of details. Further humans can process visual information better than data presented in other forms.

To implement the visual analytics framework for this project, the commonly used tools have limitations when used in as is condition, thus posing a set of challenges. There is a need for integrated statistical computation of ordinal datatype and generation of plots which is not supported by these tools. Moreover, these tools do not support customization of plots like encoding, splitting of views for comparison and grid layouts. In some of these tool, the required plots are not available and the interactions available are limited, justifying the need to build a customized dashboard.

1.3 Contribution

A visual analytics tool was developed based on Tamara Munzner four nested levels of visualization as shown in Figure 1.3 according to the needs of the researchers. The implementation lets the target users select two different population subsets of their choice and compare them to identify the similarities and differences. The visualizations developed consists of three parts based on the discussion with the researchers, the first focuses on visualizing individual symptoms to depict the distribution based on their occurrence in patients and severity. The second focuses on visualizing pairs of symptoms occurring simultaneously to identify groups and trends and the last focuses on visualizing multiple symptoms to identify patterns.

The implementation has been evaluated by conducting a user study and it was found that overall, the visualization helped in gaining new insights which is otherwise not possible due to the limited clinical knowledge in developing new hypothesis.

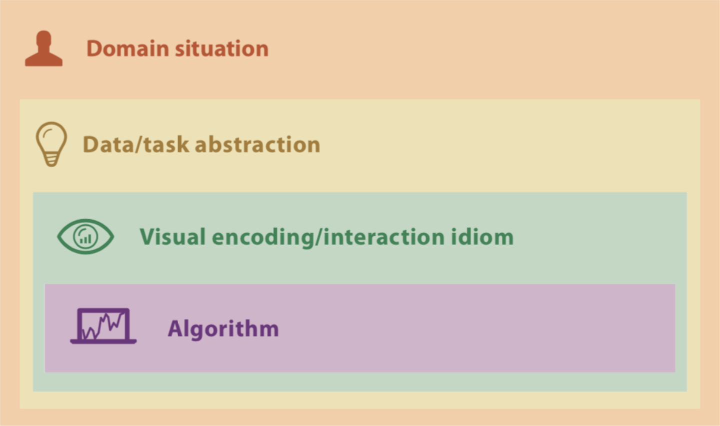


Figure 1.3: Tamara Munzner four nested levels of visualization. (Source: Visualization Analysis & Design, Tamara Munzner)

1.4 Report Organisation

Chapter 2 discusses literature in the line of visualizing patients medical records and Chapter 3 covers domain analysis. Chapter 4 discusses task and data abstraction, followed by Chapter 5 that discusses visual encoding, interaction and implementation details. Chapter 6 talks about evaluating the implementation. Chapter 7 concludes the work done and discusses potential future work.

Chapter 2

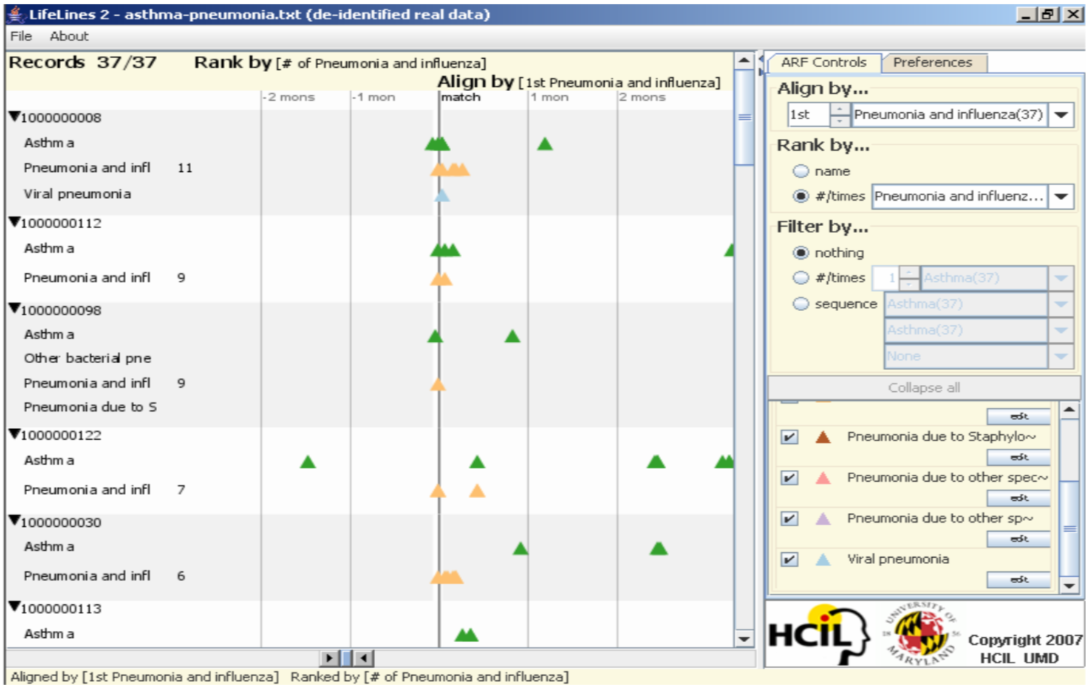
Related Work

When it comes to examining patients' medical records, in terms of visual analytics there are two focus areas according to literature [8]. There are visual analytics tools focusing on presenting medical information for individual patients and there are visual analytics tools presenting multiple patient records in parallel to help identify trends and patterns by clinicians [8]. Considering the focus of this project we focus on the later, and the literature that fall under this category are as follows.

Patient Record Intelligent Monitoring and Analysis (PRIMA), is an application that helps in visualizing and understanding patient record data [9]. It comprises of three views namely aggregate table, histogram stack and Kaplan-Meier survivability curve. These views are designed for specific data types and in case an appropriate data is not available for a view it is not displayed. The aggregate table view, shows the proportion of patients in each category for a particular variable selected and is suitable for categorical variables. The next view is called a histogram stack, which comprises of histogram for a set of selected variables and is suitable for numerical data. The last view is called the Kaplan-Meier survivability curve, which represents the probability of survival of potentially incomplete data through a process called censoring [9]. PRIMA contributes to the thesis by highlighting the visualizations possible with numerical and categorical data and how multiple views and interactive visualization can be used to identify patterns and relationships. However, the drawbacks with this visualization is that the different views are not linked, in the sense every plot has got its own selection mechanism and it is not possible to see how having a common selection mechanism will affect the visualization in different views.

InfoZoom, is a novel technique to display database contents in the form of extremely compressed tables, that ensures the size of the dataset does not lead to loss of information while visualizing due to screen size issues. Moreover, the tables not only serve visualization purpose but also lets the user perform database queries by direct manipulation of the presented information [10]. InfoZoom starts with a compact representation of the dataset and for each attribute it is possible to see the values and frequencies, and the same information is used to find the correlation between attributes. Therefore, it focuses on letting the user explore the data themselves to know the information contained in the dataset instead of using algorithms to depict information, along with how visualization can be performed on a large dataset and detect interesting knowledge without any loss of information due to screen size issues. The drawbacks associated with this visualization is that while it lets the user filter data and see how the visualization changes, it does not support comparison of different subsets of the dataset. Further, correlation is with regard to the different levels present in the attribute and does not represent the correlation between two attributes to understand the relation between them.

Lifeline2 helps in visualizing categorical temporal data by letting users of the application select records from multiple patients in the form of a query [11][12]. Each record occupies a row and is identified by its id on an absolute timescale and every medical condition has a color associated with it. Lifeline2 also lets the user to align, rank and filter the display wherein it is believed that with alignment it is easier to explore the data for potential temporal patterns across the multiple records selected. Moreover, alignment lets to compare relative timestamps easily and ranking and filtering tend to compliment alignment by reordering and narrowing the set of records interactively to suit a user's needs as shown in Figure. 2.1. This makes identifying precursor, aftereffects and co-occurring events a smooth process and serves as a good example to show how order plays an important role in visualization which has been incorporated in this thesis. This tool misses the correlation between the different medical conditions and filtering is possible only based on the presence of medical conditions. It does not help compare the results for different subsets of the dataset based on population details like age, gender and also does not give the opportunity to see how the results vary when aligned based on different medical conditions simultaneously.



LifeLines 2 - asthma-pneumonia.txt (de-identified real data)

Records 37/37 Rank by [# of Pneumonia and influenza]

Align by [1st Pneumonia and influenza]

▼1000000008
Asthm a
Pneumonia and infl 11
Viral pneumonia

▼1000000112
Asthm a
Pneumonia and infl 9

▼1000000098
Asthm a
Other bacterial pne
Pneumonia and infl 9
Pneumonia due to S

▼1000000122
Asthm a
Pneumonia and infl 7

▼1000000030
Asthm a
Pneumonia and infl 6

▼1000000113
Asthm a

Align by...
1st Pneumonia and influenza(37)

Rank by...
○ name
● #/times Pneumonia and influenz...

Filter by...
● nothing
○ #/times 1 Asthma(37)
○ sequence Asthma(37)
Asthma(37)
None

☑ Pneumonia due to Staphylo~
☑ Pneumonia due to other spec~
☑ Pneumonia due to other sp~
☑ Viral pneumonia

Aligned by [1st Pneumonia and influenza] Ranked by [# of Pneumonia and influenza]

Figure 2.1: LifeLines2 - all patients records are aligned based on the 1st reported pneumonia and influenza on a relative scale. (Source: Aligning Temporal Data by Sentinel Events: Discovering Patterns in Electronic Health Records)

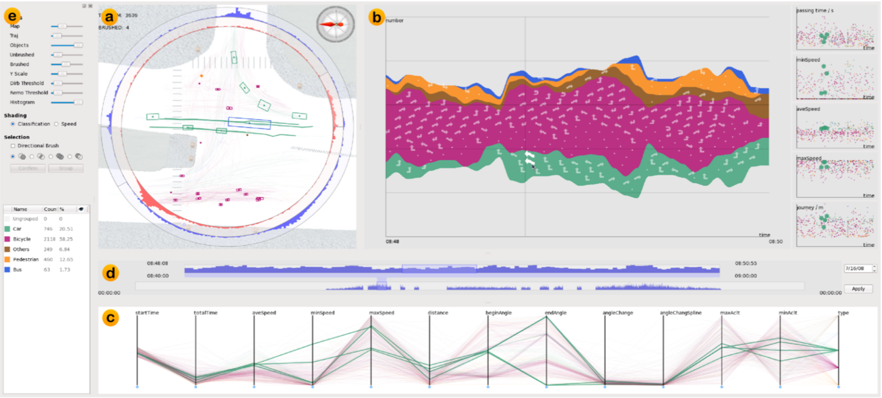
Outflow is a novel visual design that combines multiple patient records into a graph based visual representation [13]. It helps in analyzing how diseases evolve over time and connect the pathways to the outcomes of the corresponding patients to help clinicians understand how certain pathways lead to certain outcomes. In outflow, first an alignment point is selected and a directed acyclic graph is formed using all the records that satisfy the alignment point. Each edge captures symptom transition and is annotated with aggregate patient statistics. Moreover, the aggregate information affects the graphs edge width there by making it easier to understand the population contributing to each edge. This graph, captures event paths that lead to the alignment point and all event paths that occur after the alignment point [13]. Outflow, differs from the other literature's by considering a set of events that are of interest using an alignment condition and then sees how the events that occur after the alignment condition are related to the ones that meet the alignment condition, which helps in identifying trends. Similar to the visual analytics framework proposed for the thesis, in outflow the focus is on the visualization of multiple symptoms and gives its users the possibility to filter data. Here, the focus is only on multiple symptoms and does not look into the symptom spectrum in finer levels like individual and pairs of symptoms. While it is possible to filter the data, forming multiple subsets and comparison is not possible and this shortcoming has been addressed in the thesis.

Knowledge-based Navigation of Abstractions for Visualization and Exploration-II (KNAVE-II) is a knowledge based computational framework used to visualize, interpret and explore time based clinical data [8][14][15]. It supports the formulation of temporal queries, using medical domain ontology and lets users to interactively explore the results [8][14]. The interface gives user the flexibility to either visualize single patient's record or records of multiple patients [8][14][15]. The user interface consists of a domain knowledge browser on the left hand side, wherein the user can enter a query. The search table on the bottom left hand side helps find all related concepts and on the right side, raw data and their abstractions are displayed [14][15][16][17]. KNAVE-II also lets its user select their preferred temporal granularity [14][15][16][17]. KNAVE-II contributes to the thesis by showing how one can filter the desired data and have an overview of each attribute. This does not help in forming and comparing the results for different subsets nor does it help in seeing the correlation between attributes and patterns if present.

Sonja Zillner, Tamás Hauer, Dmitry Rogulin, Alexey Tsymbal, Martin Huber and Tony Solomondies propose a semantic visualization of patient information where, patient data is mapped to relevant fragments of ontologies and inferred ontologies in order to provide improved visualizations [18]. The visualization used are called semantic facet browser and semantic treemap visualization. Ontology based facet browsing enables browsing over the set of patient records and helps identify similar patient records with regard to a specific set of attributes. Treemaps on the otherhand help in identifying correlations among clinical attributes and each rectangle in the treemap corresponds to a patient [18]. Therefore, this paper helps in understanding as to how comparison and correlation play an important role in understanding the data in hand better and has been incorporated in the thesis. It is also seen that correlation helps in testing hypothesis in an efficient manner. However, the short comings associated are as follows, semantic facet browser displays the information in tabular format which makes it difficult to identify the differences. Similarly, treemaps are used for correlation and as the number of records increase, it will be difficult to infer from the treemap as it gets cluttered.

Lifeflow is a novel interactive visual overview of event sequences, that is scalable and can summarize all possible sequences. It also represents the temporal spacing of the events within sequences and is an extension of the Lifeline2 [11][12]. The working of Lifeflow is as follows, raw data is represented as colored triangles on a horizontal timeline and each row represents a record. The records are aggregated by sequence into a data structure called tree of sequences. The tree of sequences is then converted into a Lifeflow visualization. Each tree node is represented with an event bar, where the height of the bar is proportional to the number of records and the horizontal position is determined by the average time between events [19]. Lifeflow, highlights the importance of depicting overview information as it lets you see the trends specific to the entire dataset easily and this concept is applied in this thesis for visualizing an overview of relevant attributes. Similar to Lifeline2, it does not help compare the results for different subsets of the dataset based on population details like age, gender and also does not give the opportunity to see how the results vary when aligned based on different values simultaneously.

There are other domains that have similar structure of dataset in comparison to the patient record dataset. An example of such a domain is traffic data, which is temporal in nature and involves different types of data like spatial, numerical and categorical [20][21]. TripVista, is a visual analytics tool built to visualize traffic data and it consists of a spatial view for depicting geometrical trajectory information, a temporal view of themeriver and scatterplots [21]. The next view involves a parallel coordinate plot to depict information from multiple attributes, a time slider to select a two-level time range selection and lastly a control panel for system parameter settings and data classification [21]. As shown in Figure. 2.2 the visualization has three views namely the spatial, temporal and multidimensional. This visualization helps in understanding how to place different views, provide interactions and system parameter selection like rendering transparency, scale of histogram along with showing how multidimensional data are used which is of particular interest. The user is not given the option to select different subsets of data and considers only multiple attributes when individual and pairs of attributes could also be considered, which has been addressed in this thesis.



Shading
● Classification ○ Speed

Selection
□ Directional Brush

	Name	Coun	%
	Ungrouped	0	0
🟩	Car	746	20.51
🟪	Bicycle	2118	58.25
🟫	Others	249	6.84
🟧	Pedestrian	460	12.65
🟦	Bus	63	1.73

Figure 2.2: Interface of TripVista. (Source: TripVista: Triple Perspective Visual Trajectory Analytics and Its Application on Microscopic Traffic Data at a Road Intersection)

From the above literature review the following conclusions can be drawn, majority of the visualizations focus on depicting the data on a timeline based on which trends and patterns are identified. Interaction is used to zoom in on a particular portion of the timeline or used to display patient details like age, gender when opposed to interacting with a plot and reflecting the changes in the other plots present or to make selections. Apart from timeline, other visualizations involve aggregation and correlation using the data as such without using any correlation techniques. This direct method of identifying correlation is not an ideal way, as when the dataset grows in size it gets difficult to infer. Overall, it can be seen that the above literature's do not focus on showing the distributions of attributes, correlation between them using computational methods, identifying relationships, finding patterns, clusters and to compare all the results for different subsets in a single visual analytics tool.

Chapter 3

Domain Analysis

The project has got a narrow target group, comprising of researches from the sleep medicine center, Kempenhaeghe. Narcolepsy is a chronic sleep disorder, that is currently identified and treated by a set of few symptoms. It is known that the symptom spectrum is not limited to the few commonly occurring set of symptoms. Hence, the researches at sleep medicine center, Kempenhaeghe decided to analyze a list of twenty symptoms related to narcolepsy with the help of the data collected through a mobile application.

After a series of discussions with the researchers, it was understood that their intention behind building the mobile application involves two goals. The first is, understanding the symptom spectrum on the whole and the second is, to give personalized treatment to patients based on their reporting and medical history. The data collected has a time component associated with it, since a user can log symptoms multiple number of times both within a day and in the subsequent days.

It was mutually agreed with the researchers that this project would focus on building visualization to understand the population and in identifying relationships between symptoms for different population subsets independent of time.

The dataset at hand is collected through the mobile application built and consists of a questionnaire followed by an interface that allows the user to select the symptoms they experience and rank them according to the severity level at the time of reporting.

The questionnaire is filled by patients only once when they first log into the application and, includes questions related to general information about the patient along with questions related to how often they experience a few symptoms. The exact attributes involved with the questionnaire are shown in Table 3.1. The primary reason behind asking the users to report how often they experience these symptoms is to compute the Ullanlinna score as a score of 14 and above is representative of narcolepsy. To compute this score the symptoms need to be reported based on frequency and not severity.

General Attributes	Symptoms
Patient Id	Knees unlocking
Name	Mouth Opening
Age	Head nodding
Gender	Falling down
Diagnosed by physician	How fast the patient usually falls asleep in the evening
Country	How often do they sleep during the day
Email	Do they fall asleep while reading
Education completed	Do they fall asleep while travelling
Employment status	Do they fall asleep while standing
	Do they fall asleep while eating
	Do they fall asleep in unusual situations

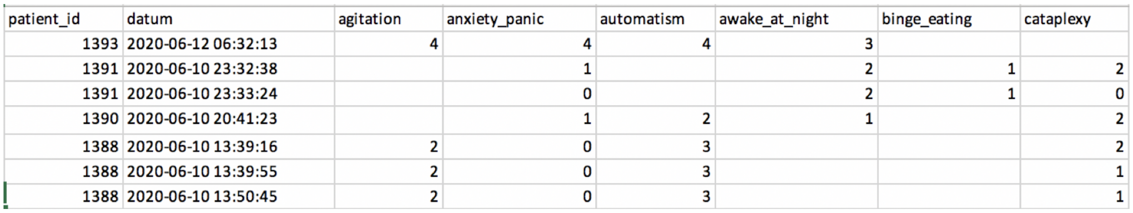
Table 3.1: Table depicting the list of attributes and symptoms collected via the questionnaire interface

The patients then get to report the symptoms they are experiencing and rank them for severity. A symptom severity can be ranked from 0 to 4 where 0 means the symptom is mild and 4 represents the symptom is extremely severe. The twenty symptoms from which the user can choose and rank are shown in Table 3.2.

Symptoms			
agitation	cataplexy	lack of energy	problems work
anxiety panic	difficulty achieving	lifelike dreams	sadness
automatism	difficulty concentrating	problems libido	sleep paralysis
awake at night	difficulty memory	problems relationship	sleepiness day
binge eating	increase weight	problems school	social contact

Table 3.2: Table depicts the list of 20 symptoms from which patients can select and rank the symptoms experienced.

Figure 3.1 gives a snapshot of patients reporting with regard to symptoms agitation, anxiety panic, automatism, awake at night, binge eating and cataplexy.



patient_id	datum	agitation	anxiety_panic	automatism	awake_at_night	binge_eating	cataplexy
1393	2020-06-12 06:32:13	4	4	4	3		
1391	2020-06-10 23:32:38		1		2	1	2
1391	2020-06-10 23:33:24		0		2	1	0
1390	2020-06-10 20:41:23		1	2	1		2
1388	2020-06-10 13:39:16	2	0	3			2
1388	2020-06-10 13:39:55	2	0	3			1
1388	2020-06-10 13:50:45	2	0	3			1

Figure 3.1: Snapshot of the dataset

Every time a patient reports, the date and time of reporting along with the medication taken and any specific note left by the patient is made available through the dataset. As can be seen from Figure. 3.1, the blanks represents symptoms not reported and should not be misinterpreted as a missing value. Therefore, when visualizing only records reported by patients are taken into consideration.

After discussions with the users, the attributes not taken into consideration are name, country, email, education completed, employment status and medication. Name and email are patient specific information and do not contribute towards forming interesting subsets. Since the project looks at the symptom spectrum irrespective of location, education and employment, the corresponding attributes are not taken into account. Lastly, medication varies with the reporting of patients which is dependent on time and hence it is not considered.

The attributes considered are, age, gender, diagnosed by physician, symptoms reported by the patients and, the date and time reported. To avoid over representation of specific situation and patients, the records need to have a certain time between them.

With the domain interest and data in hand, the tasks for exploration and analysis of the data are the following,

1. Filter records and select interesting subsets of the dataset based on relevant attributes like age, gender and diagnosed by physician.
2. Analyze and explore distribution of individual symptom -
   (a) Distribution of occurrence of symptoms on population subset.
   (b) Distribution and trends on severity of individual symptoms reported.
3. Analyze and explore relation between pair of symptoms -
   (a) Are there two symptoms that are similar to each other based on whether they occur simultaneously?
   (b) Identify trends and groups between pairs of symptoms.
4. Analyse and explore relationships between multiple symptoms -
   (a) Identify patterns involving multiple symptoms at a time.
5. Be able to compare symptom characteristics based on subsets of the dataset using relevant attributes like age, gender and diagnosed by physician.

Chapter 4

Task and Data Abstraction

4.1 Data Abstraction

The dataset at hand is in the form of a flat table where each row represents an item of the data and each column represents an attribute.

The dataset used for the visualization is available all at once and is a static dataset file. Information required is downloaded from the mobile application server and new information cannot flow during the course of visualization.

The data types involved in this dataset are categorical, date time, identifiers and ordered. Under ordered both ordinal and quantitative data types are present. Table. 4.1 gives an overview of the attributes falling under each data type.

Categorical	Date time	Identifier	Ordinal	Quantitative
Gender Diagnosed by physician	data and time	Patient id	Symptoms including those in the questionnaire and the 20 symptoms for which patients report severity	Age

Table 4.1: Table briefly depicts the attributes falling under each data type

Ordered data are generally associated with a direction and in this case the ordering direction is from 0 to 4 representing mild to extremely severe. Our focus is on these ordinal attributes.

4.2 Task Abstraction

After having looked into the domain analysis, the task abstraction for the project is as follows,

Tamara Munzner has defined 3 levels of actions that helps define user goals and they are explained below.

- The project focuses on helping the target users consume information by letting them **discover** and analyze new information.
- The project helps **explore** trends, relations and groups.
- The project focuses on **comparing** the results obtained for different targets. Target corresponds to the aspects of data that is of interest to the users.

As shown in the domain task specification in Chapter 3 it would be of interest to look into one, two and multiple attributes at a time. This implies that the targets here are attributes and the abstract tasks that need to be carried out are,

1. Filtering and selection.
2. When considering individual attributes -
   (a) The task is to identify distribution of categorical attributes.
   (b) The task is to identify distribution of ordinal attributes and trends.
3. When considering two attributes at a time -
   (a) The task is to be able to identify pairs of attributes that occurs simultaneously. In order to achieve the same, association and agreement between every pair of attributes needs to be computed.
   (b) Identifying groups and trends in two attributes.
4. When considering multiple attributes at a time -
   (a) The task is to be able to identify patterns between multiple attributes.
5. Compare results for different selections made.

Chapter 5

Visual encoding and Implementation

This chapter introduces visual encoding and interactions designed from the task analysis. We also present the tool used for implementation and present the potential limitations.

5.1 Tool used for Implementation

This project is implemented using R which is an open source programming language widely used for statistical computing and graphics [1]. The primary reason of opting for R over other programming tools is due to the fact the project required a tool capable of supporting statistical computation like association and to build an interactive visualization dashboard. R is supported by a number of operating systems including UNIX-like, Windows and MacOS.

The functionalities of R can be extended with the help of packages and this project makes use of two main packages namely ggplot2 [2] and shiny [3] for plotting the visualization and for interaction respectively.

5.2 Overview

Figure. 5.1 gives an overview of the tasks, details of visual encoding and interactions which are covered in detail in the following sections.

[1] https://www.r-project.org/about.html

[2] https://ggplot2.tidyverse.org/

[3] https://shiny.rstudio.com/

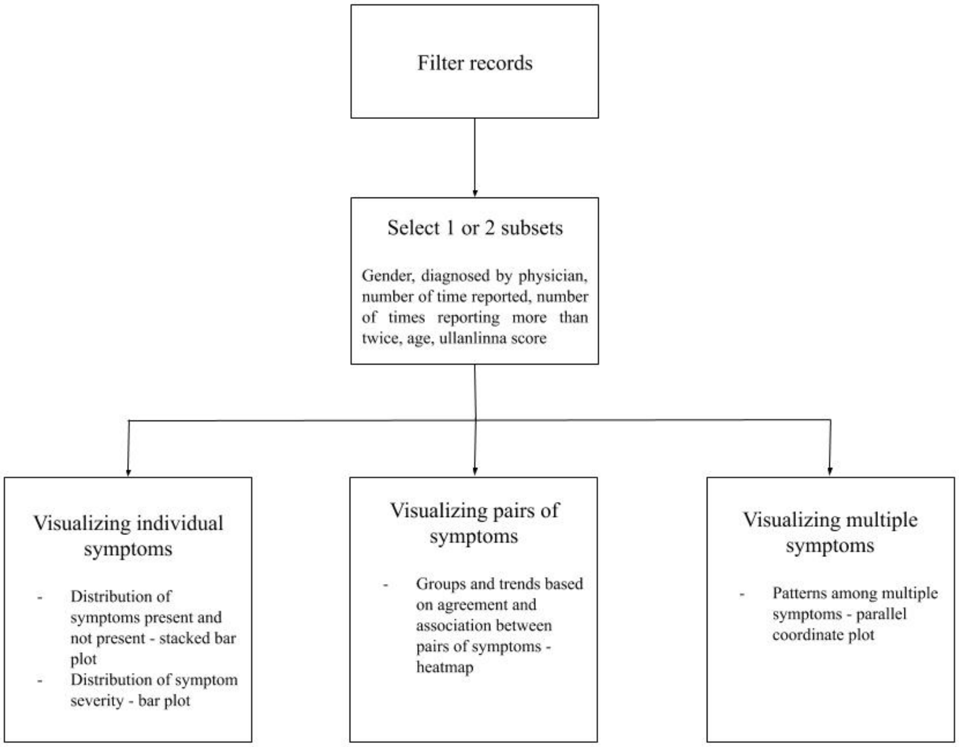

Filter records
Select 1 or 2 subsets
Gender, diagnosed by physician, number of time reported, number of times reporting more than twice, age, ullanlinna score
Visualizing individual symptoms
- Distribution of symptoms present and not present - stacked bar plot
- Distribution of symptom severity - bar plot
Visualizing pairs of symptoms
- Groups and trends based on agreement and association between pairs of symptoms - heatmap
Visualizing multiple symptoms
- Patterns among multiple symptoms - parallel coordinate plot

Figure 5.1: Overview of tasks

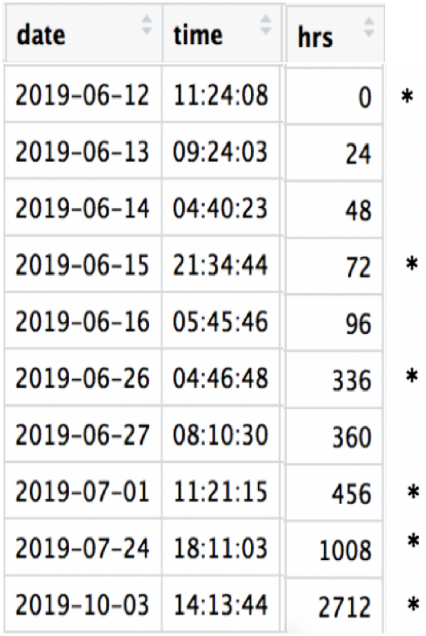
5.3 Filtering - Subset Selection

This section corresponds to task 1 in Chapter 4 and it involves two steps. The first is filtering of records to avoid data bias (Subsection 5.3.1) and the second is definition of population subsets for further analysis (Subsection 5.3.2).

5.3.1 Filtering of records

Here we focus on the relation between symptoms independent of the time component. However, a patient can report multiple times both within a day and on different days, in such a case the records considered might bias the analysis. Hence the users were given the option to either select the first record reported by every patient or to select the first record reported by every patient along with records which are at least certain timestamp apart with regard to the previous record. To avoid over representation of a patient, we allow the user to select the required time between records ensuring only one record per day is selected.

This is illustrated as shown in Figure. 5.2 for a patient when the timestamp chosen between records is 50 hours.



date	time	hrs	
2019-06-12	11:24:08	0	*
2019-06-13	09:24:03	24	
2019-06-14	04:40:23	48	
2019-06-15	21:34:44	72	*
2019-06-16	05:45:46	96	
2019-06-26	04:46:48	336	*
2019-06-27	08:10:30	360	
2019-07-01	11:21:15	456	*
2019-07-24	18:11:03	1008	*
2019-10-03	14:13:44	2712	*

Figure 5.2: Filtering of records when timestamp entered is 50 hours

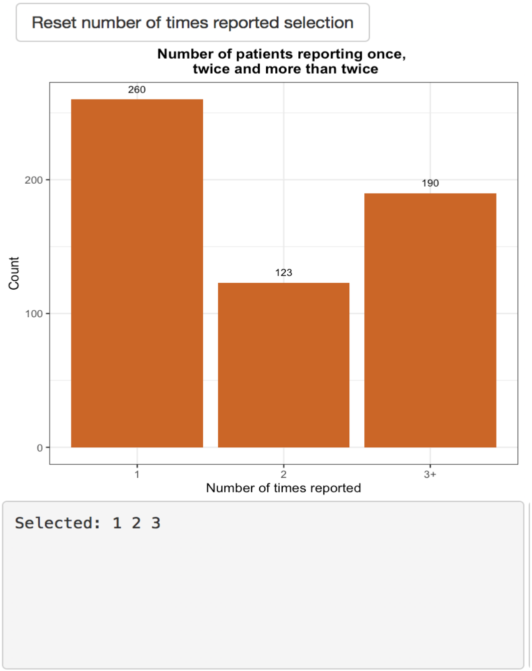
5.3.2 Subset selection

The filtering mechanism provided to the target users to form interesting population subset is as follows. The first attribute used to form population subset is *diagnosed by physician* followed by *gender*. Under diagnosed by physician, the target user can select based on whether a patient is diagnosed or not. A similar selection option is available for gender.

Another attribute for selection is the *number of times a patient has reported*. This attribute gives the target users a clear picture as to how often the users of the dataset are actually using the app.

To visually depict the distribution of patient reporting we use bar plots as shown in Figure. 5.3. Bar plots are known to be the most accurate to visualize two attributes which are ordered and quantitative [22].

The selection is effective by clicking on the bars. As can be seen from Figure. 5.3 the selection made is printed in a textbox and the frequency of each category is printed on top of the corresponding bars.

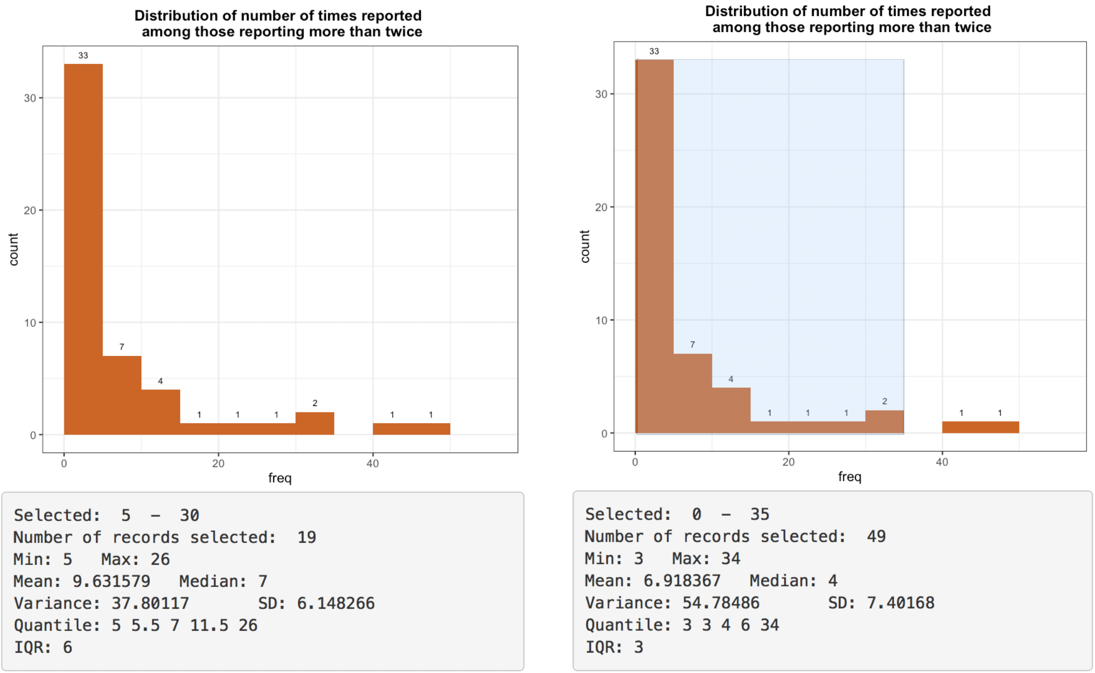

Reset number of times reported selection
Number of patients reporting once, twice and more than twice
260
123
190
Count
0
100
200
1
2
3+
Number of times reported

```
Selected: 1 2 3
```

Figure 5.3: Distribution of number of times reported by patients using the mobile application

For more detailed information it is also possible to use a histogram to depict the different *number of times patients report among those reporting more than twice*. Since there are 2 continuous attributes and the goal here is to look into the distribution a histogram is used as shown in Figure. 5.4a with a bin width of 5 [22].

As can be seen from Figure. 5.4a value is printed above each bin and the target user can select the preferred range by brushing the histogram as shown in Figure. 5.4b. A text box is used to display information like the mean, median, minimum value, maximum value, quantile, IQR and the number of records involved in the selection.



```
Selected:  5  –  30
Number of records selected:  19
Min: 5   Max: 26
Mean: 9.631579   Median: 7
Variance: 37.80117       SD: 6.148266
Quantile: 5 5.5 7 11.5 26
IQR: 6
```

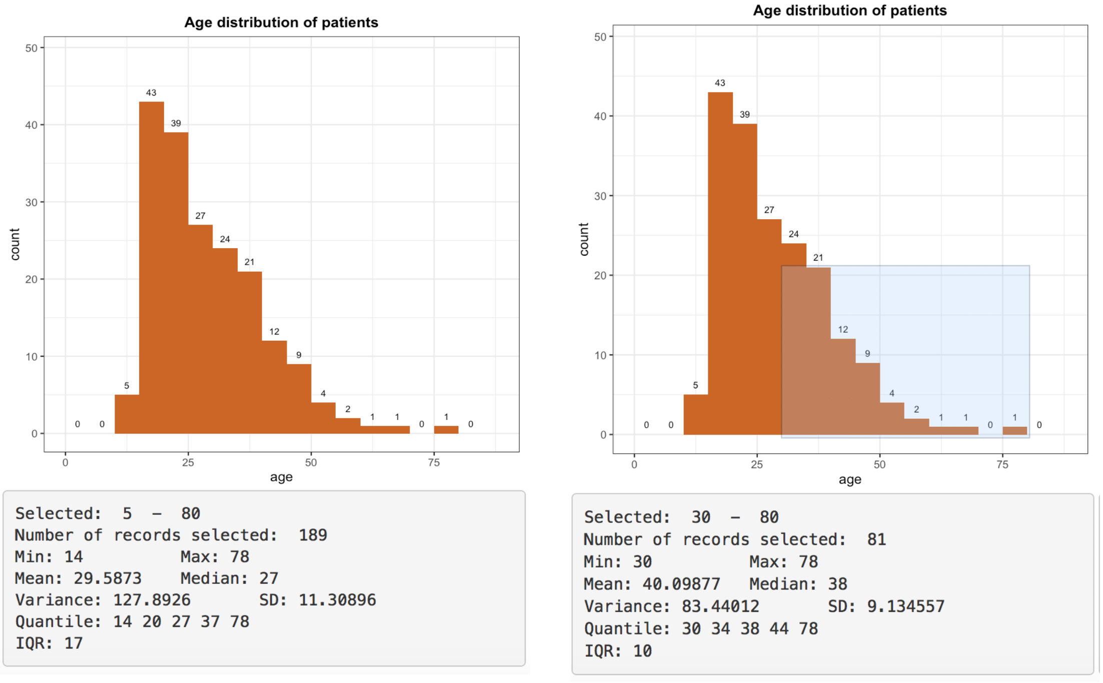
```
Selected:  0  –  35
Number of records selected:  49
Min: 3   Max: 34
Mean: 6.918367   Median: 4
Variance: 54.78486       SD: 7.40168
Quantile: 3 3 4 6 34
IQR: 3
```

(a) Distribution plot

(b) Interaction - brushing

Figure 5.4: Number of times reported among those reporting more than twice

The next attribute given to the target user to select the population subset of interest is *age*. Here again a histogram is used as shown in Figure. 5.5a with interaction provided in the form of brushing as shown in Figure. 5.5b.



```
Selected:  5  –  80
Number of records selected:  189
Min: 14          Max: 78
Mean: 29.5873    Median: 27
Variance: 127.8926      SD: 11.30896
Quantile: 14 20 27 37 78
IQR: 17
```

```
Selected:  30  –  80
Number of records selected:  81
Min: 30          Max: 78
Mean: 40.09877   Median: 38
Variance: 83.44012      SD: 9.134557
Quantile: 30 34 38 44 78
IQR: 10
```

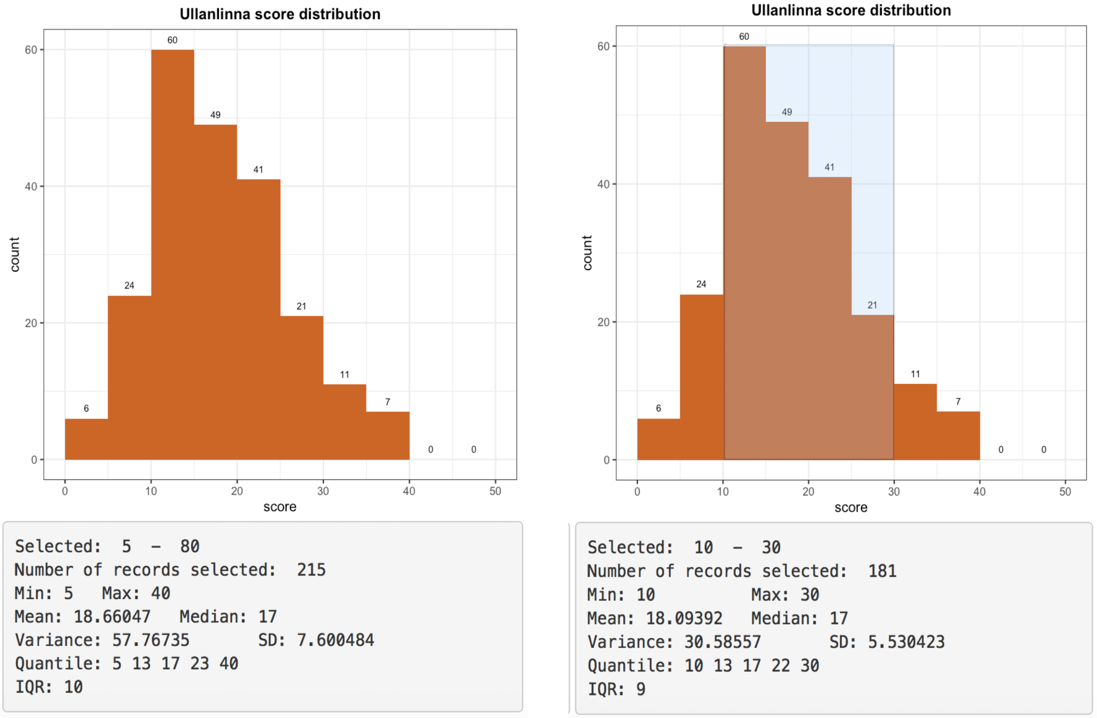
(a) Distribution plot

(b) Interaction - brushing

Figure 5.5: Age

The last attribute given to the target users to form a population subset of their choice is *Ullanlinna score*. This score is computed for each patient by assigning points for every symptom in the questionnaire [7]. The points assigned are then added and the value obtained is the Ullanlinna score.

Here again a histogram is used as shown in Figure. 5.6a and interaction is provided in the form of brushing as shown in Figure. 5.6b



```
Selected:  5  –  80
Number of records selected:  215
Min: 5   Max: 40
Mean: 18.66047   Median: 17
Variance: 57.76735       SD: 7.600484
Quantile: 5 13 17 23 40
IQR: 10
```

```
Selected:  10  –  30
Number of records selected:  181
Min: 10          Max: 30
Mean: 18.09392   Median: 17
Variance: 30.58557       SD: 5.530423
Quantile: 10 13 17 22 30
IQR: 9
```

(a) Distribution plot

(b) Interaction - brushing

Figure 5.6: Ullanlinna score

This selection mechanism apart from helping in forming interesting subsets also serves as a means to understand the dataset population better.

5.4 Visualization of individual attributes

In this section we cover two subtasks related to visualizing individual attributes. The first is visualizing the distribution of symptoms present and not present (Subsection. 5.4.1). The second is visualizing distribution of symptom severity and identifying trends (Subsection. 5.4.2).

5.4.1 Distribution of symptoms present and not present

In this subsection we look into the distribution of individual symptoms present and not present. This corresponds to task 2(a) in Chapter 4. A stacked bar is used as shown in Figure. 5.7 as it is more appropriate in understanding the presence or absence of symptom with respect to the total number of records in the population subset [22].

Color is selected based on Colorbrewer [4] for categorical attribute with two categories namely present and not present.

[4] https://colorbrewer2.org

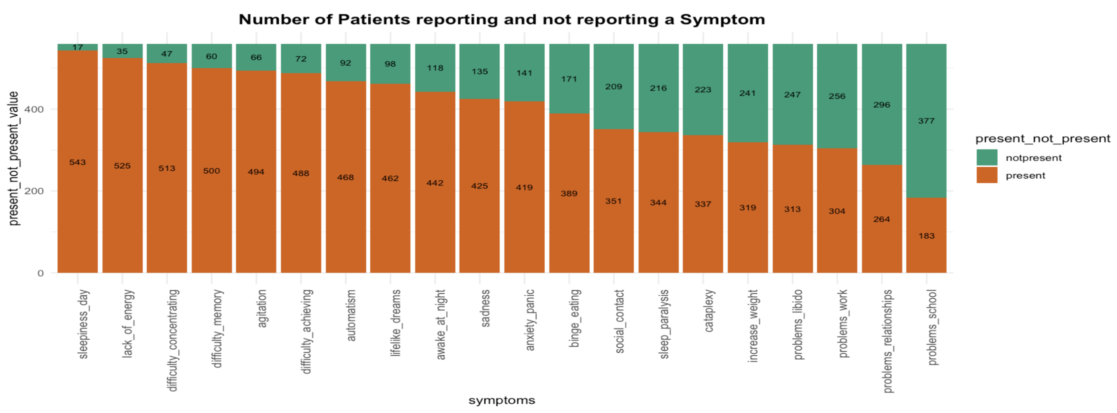


Figure 5.7: Stacked bar plot depicting the distribution of symptoms in terms of those reporting and not reporting a symptom

Since seeing the patients in whom a symptom is present is of importance, the same is chosen to order the stacked bar plot. Ordering the stacked bar plot helps in seeing patterns better [22]. In every stack the exact frequency value is printed.

5.4.2 Distribution of symptom severity and identification of trends

The next task is to visualize the distribution of symptoms based on severity. This corresponds to task 2(b) in Chapter 4. A bar plot is used as shown in Figure. 5.8 as it is considered to be accurate in visualizing two attributes which are ordered and quantitative [22]. The bar plots pertaining to each individual symptom are displayed all at once in a 4 cross 5 grid to enable better comparison.

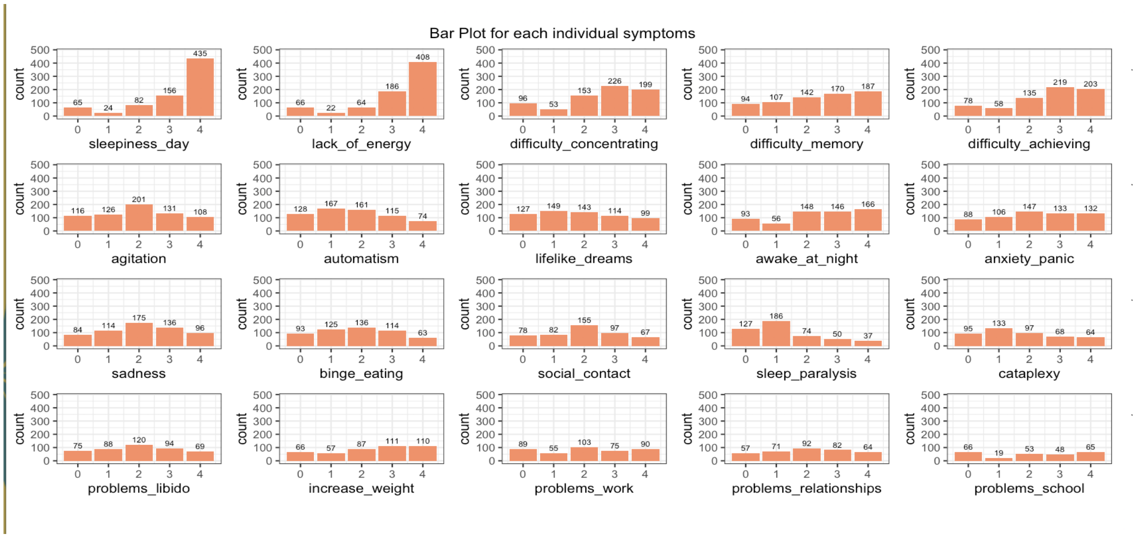


Figure 5.8: Individual bar plots depicting the distribution of symptoms in terms of severity

These bar plots are ordered according to the order of their corresponding stacked bar plot and when these are ordered, the changes are reflected on the individual bar plots. Similar to the stacked bar plots, the frequency value of each bar is printed on top.

5.5 Visualization of pairs of attributes

This section covers two subtasks pertaining to task 3 in Chapter 4. The first is identifying pair of attributes that occur simultaneously based on reporting and severity (Subsection 5.5.1, Subsection 5.5.2). The second is identifying groups and trends (Subsection 5.5.3).

5.5.1 Co-occurrence of symptoms based on reporting

In this section, we cover visualization of two symptoms that occur simultaneously based on reported and not reported. This corresponds to task 3(a) in Chapter 4. A contingency table which shows the relationship between two categorical attribute is used for this purpose [5]. This is visually represented using a balloon plot as shown in Figure. 5.9.

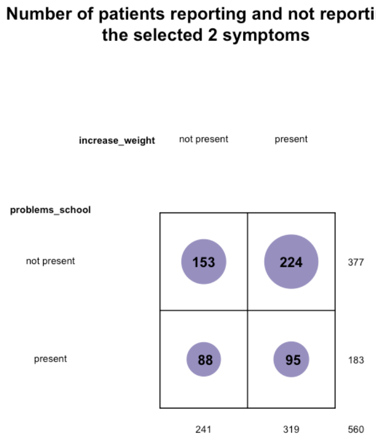


Figure 5.9: Balloon plot depicting the number of patients reporting different combinations of reporting and not reporting

The diagonal in Figure. 5.9 represents the agreement between two symptoms. Agreement is defined as the relation between pair of symptoms based on whether they are reported or not and is computed as follows,

$$A = N1 + N2 \tag{5.1}$$

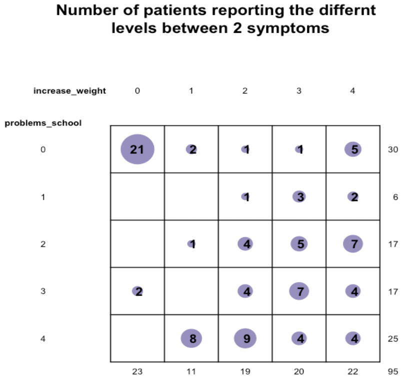
where,

A = agreement
N1 = number of patients reporting both the symptoms
N2 = number of patients not reporting both the symptoms

[5] https://en.wikipedia.org/wiki/Confusion$_{m}atrix$

5.5.2 Relation of symptoms based on severity

In this section, we cover visualization of two attributes at time to identify symptoms occurring simultaneously based on severity and this also corresponds to task 3(a) in Chapter 4. Here again, contingency table is used to represent the number of times patients report the different combinations of severity present in a pair of symptoms. This is visually represented using a balloon plot as shown in Figure. 5.10.



problems_school \ increase_weight	0	1	2	3	4	
0	21	2	1	1	5	30
1			1	3	2	6
2		1	4	5	7	17
3	2		4	7	4	17
4		8	9	4	4	25
	23	11	19	20	22	95

Figure 5.10: Balloon plot depicting the number of patients reporting different combinations of symptom severity

The balloon plot as shown in Figure. 5.10 helps in seeing the association based on the frequency distribution of severity levels. Association refers to the relation between pair of symptoms based on severity and can also be computed directly as follows,

Unlike quantitative variables or categorical variables with two categories, in this project we are dealing with ordinal attributes with 5 levels due to which commonly used correlation methods like Pearson's correlation coefficient is not an appropriate choice. Moreover, Pearson's correlation is suitable when dealing with normalized data and when looking for linear relationships. Therefore, the choice of correlation should be distribution free and should be based on ranking of symptom severity. There are two common methods that can be used namely spearman's rank correlation and Kendall's rank correlation.

In general, the spearman's correlation is considered the non-parametric counterpart of the Pearson's correlation whereas Kendall rank correlation computes the correlation based on concordant and discordant pairs. However, both spearman and Kendall are suitable for the given dataset as we are computing the association between two ordinal attributes, but in general spearman's correlation is found to be sensitive to error and is more suitable for ordinal attributes with more than six levels [23], hence Kendall rank correlation is used to find the association between pairs of symptoms.

The Kendall rank correlation is computed as follows,

$$\tau = \frac{(number\ of\ concordant\ pairs) - (number\ of discordant\ pairs)}{n(n-1)/2} \tag{5.2}$$

where,

n = *number of records*

For example let us consider patient A and patient B, each representing an item. Now if we are looking at the association between two symptoms namely agitation and anxiety the concordant and discordant pairs are defined as follows,

Concordant pairs:

$$Patient_A >_{agitation} Patient_B$$

$$Patient_A >_{anxiety} Patient_B$$

Discordant pairs:

$$Patient_A >_{agitation} Patient_B$$

$$Patient_A <_{anxiety} Patient_B$$

When two patients report the same severity level of a symptom it is called a tie. Equation 5.2 is not suitable when ties are present in the dataset. When a tie is present the two records which are being compared have the same value and cannot be ranked due to which it does not contribute to the ranked correlation. Hence, Equation 5.2 is modified as follows and is called the Kendall tau – b rank correlation [6].

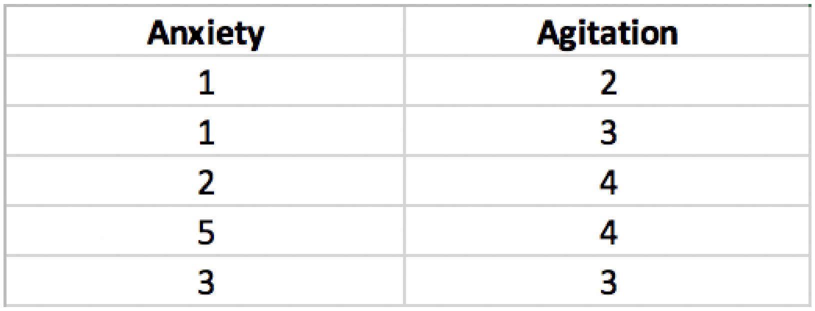
$$\tau_b = \frac{(number\ of\ concordant\ pairs) - (number\ of discordant\ pairs)}{\sqrt{N_1} \times \sqrt{N_2}} \tag{5.3}$$

where,

N_1 = *number of pairs with different severity levels for symptom 1*
N_2 = *number of pairs with different severity levels for symptom 2*

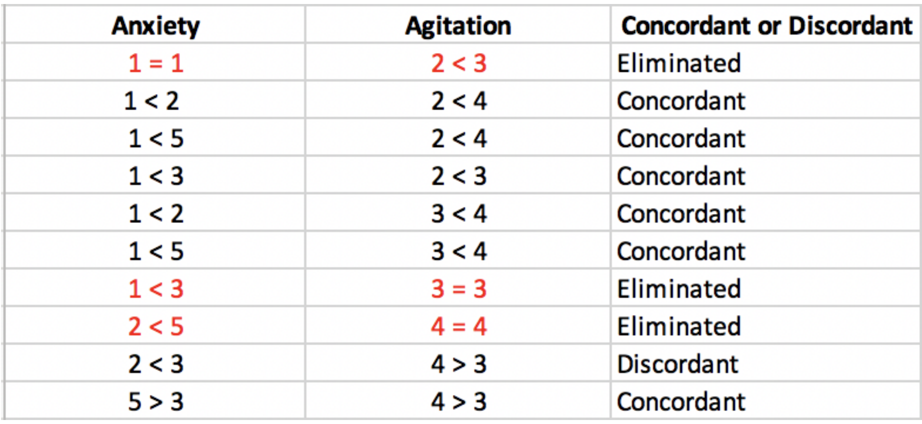
Association can range from -1 to 1 and the following is an illustration of the kendall tau-b for two symptoms, anxiety and agitation with levels 1 to 5. Figure. 5.11 is the toy dataset used and Figure. 5.12 illustrates the formation of concordant and discordant pairs.

[6] https://www.r-tutor.com/gpu-computing/correlation/kendall-tau-b



Anxiety	Agitation
1	2
1	3
2	4
5	4
3	3

Figure 5.11: Toy dataset to illustrate kendall tau b for symptoms anxiety and agitation



Anxiety	Agitation	Concordant or Discordant
1 = 1	2 < 3	Eliminated
1 < 2	2 < 4	Concordant
1 < 5	2 < 4	Concordant
1 < 3	2 < 3	Concordant
1 < 2	3 < 4	Concordant
1 < 5	3 < 4	Concordant
1 < 3	3 = 3	Eliminated
2 < 5	4 = 4	Eliminated
2 < 3	4 > 3	Discordant
5 > 3	4 > 3	Concordant

Figure 5.12: Illustration of kendall tau b for the toy dataset

Pairs with ties are eliminated and the kendall tau-b rank correlation according to Equation 5.3 is,

$$\tau_b = \frac{6-1}{\sqrt{9} \times \sqrt{8}} = 0.58$$

5.5.3 Identify groups and trends

The balloon plots helps in seeing the agreement and association between a pair of symptoms based on reporting and severity respectively. However, it is of interest to see the agreement and association between every pair of symptoms all at once for which a heatmap is used as shown in Figure. 5.13. The agreement and association values computed using Equations 5.1 and 5.3 is used to generate the heatmap.

Heatmap is a two dimensional matrix representation of the information and an alternative to the heatmap is a network visualization consisting of nodes and edges. Heatmaps rely on colors to distinguish and interpret the information, making it easier to grasp the key insights and they also help in seeing an overview of the information [24][7]. They are suitable for large datasets, clutter free and help identify patterns and relationships making it an ideal choice.

[7] https://www.dundas.com/resources/blogs/when-and-why-to-use-heat-maps

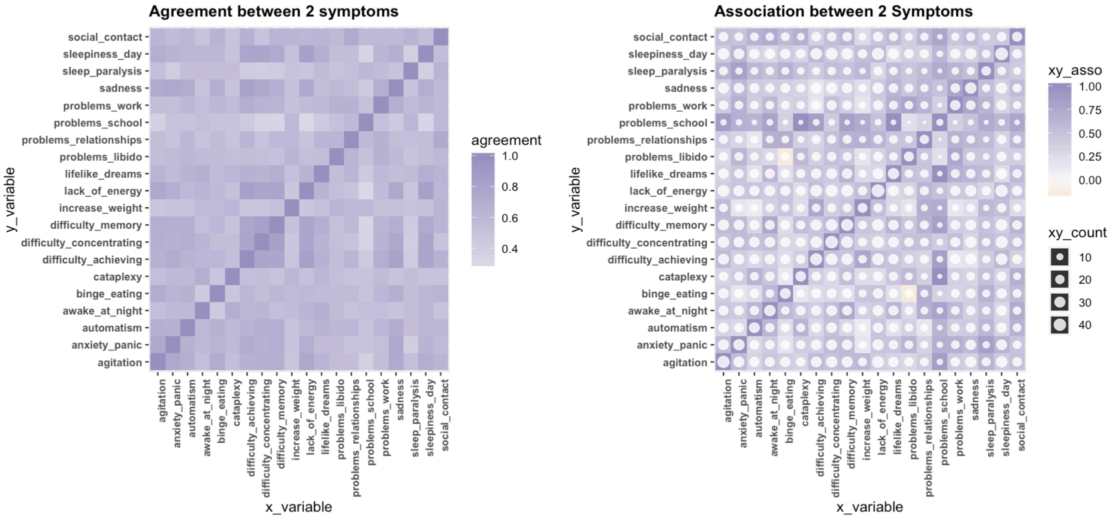


Figure 5.13: Heatmaps for agreement and association between pairs of symptoms

The views corresponding to the two population subsets selected are further split so that the association and agreement heatmap are adjacent to each other.

The color is selected using Colorbrewer [8] and as can be seen from Figure. 5.13 a diverging color palette is used. For agreement the values range between 0 to 1 as it is represented in percentage and association ranges from -1 to 1. The negative values are in the shades of orange, values close to zero are in the shades of white and positive values are in the shades of violet. In the association heatmap when only one record is available for a pair of symptom then not available (NA) is returned by kendall tau-b function and grey color is used to represent the same.

In Figure. 5.13 the circles represent the number of records used to compute the association as not all patients would have reported the different combination of pairs of symptoms. How this affects the association is important to understand and the user is given the flexibility to decide if they want to encode the heatmap with this information or not. To ensure that this does not distract the user's attention from the underlying heatmap the saturation has been reduced.

In Figure. 5.13 it is hard to infer patterns and ordering the rows and columns of the heatmap based on symptoms related to each other help identify trends and groups as per task 3(b) in Chapter 4 [24][9]. To help provide a deeper exploration of the data at hand the target users are given the choice to select the order of the heatmap based on name, value and cluster. Figure. 5.13 shows the heatmap generated when name is chosen.

Ordering the heatmap based on value takes the average of every row and column and arranges them in ascending order as shown in Figure. 5.14 . This lets the target user see symptoms that have a low or high agreement and association.

[8] http://colorbrewer2.org/

[9] https://bost.ocks.org/mike/miserables/

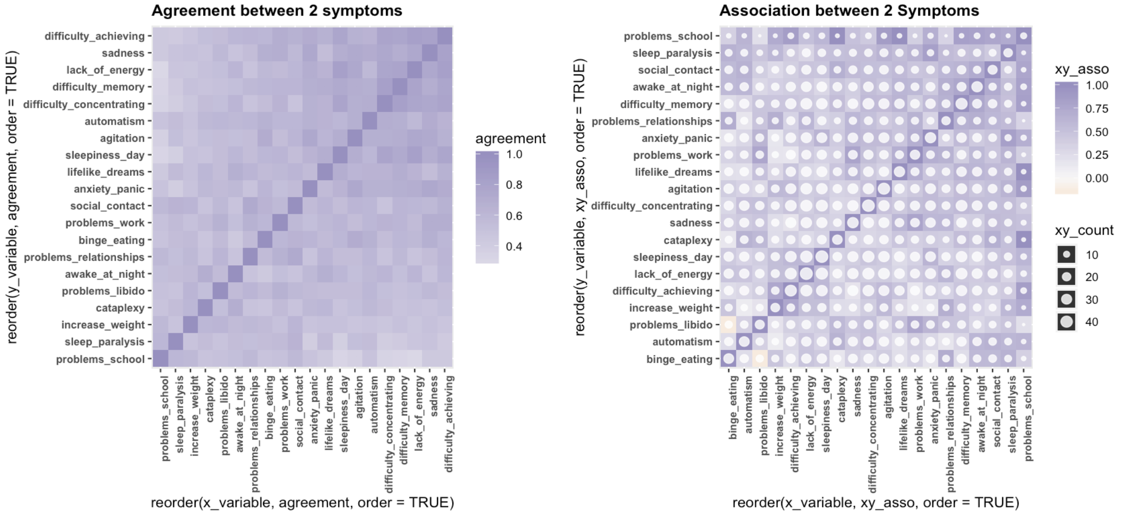


Figure 5.14: Heatmaps ordered based on value

Clustering lets the target users see if there are symptoms that can be grouped and those that form a square around the diagonal can be grouped together. For example in Figure. 5.15 in the agreement heatmap, one such cluster is sleepiness day, difficulty achieving, lack of energy, difficulty concentrating and difficulty memory.

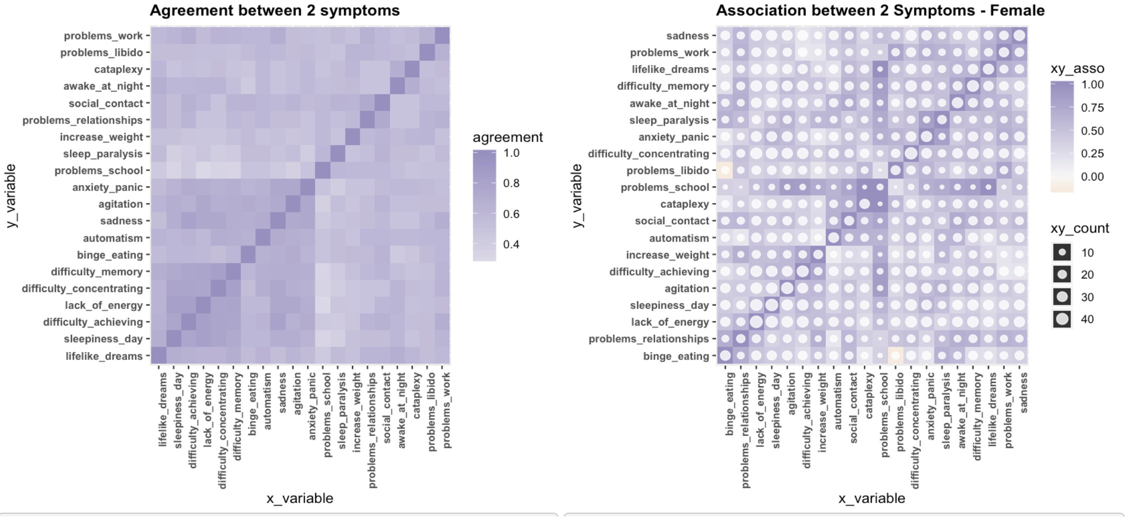


Figure 5.15: Heatmaps ordered based on cluster

Identifying groups of symptoms which are similar to each other based on reporting and severity experienced helps in forming new hypothesis which aids in better diagnosis.

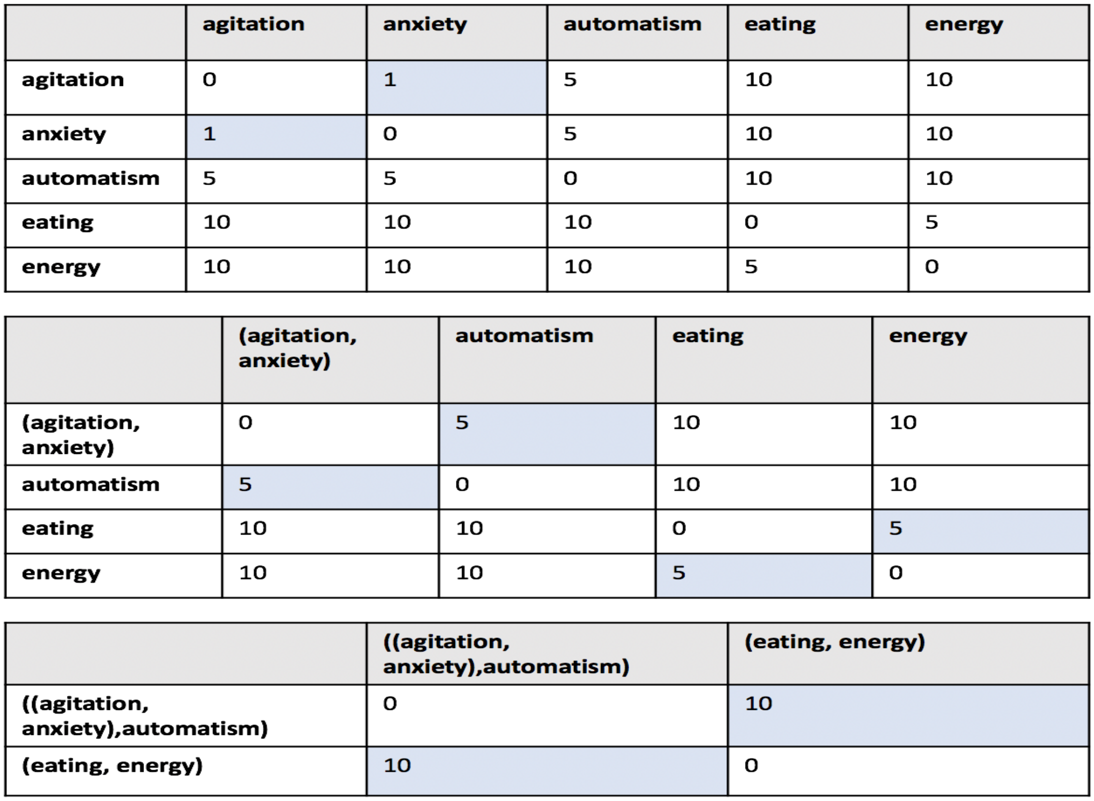
Hierarchical clustering is used where clusters are formed hierarchically using symptom similarities. It is preferred as the results obtained are reproducible, that is every time the program is run for the given dataset the clusters obtained are the same [10]. Moreover, the number of clusters required need not be known before hand. The method used to form

[10] https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/

the clusters is called agglomerative where a bottom-up approach is used. Each symptom is considered to be an independent cluster and these clusters are grouped by merging the clusters which are similar to each other. Complete linkage is the criteria used as the distance between clusters should be maximum. The reason being similar symptoms should be in a cluster and symptoms which are not similar should be far apart.

Similarity between symptoms which in this case is agreement and association are used as the distance measure directly. In doing so symptoms with smaller distance are clustered together, that is symptoms with low similarity between them are grouped. However, it is desired that symptoms with higher similarities are grouped together and for this purpose the similarity needs to be inverted. Agreement ranges from 0 to 1 and to invert we have to take 1 – the agreement value. Association ranges from -1 to 1 and we invert it such that the scale ranges from 0 to 1. It is inverted using the formula 1 – ((association + 1)/2). The inverted agreement and association values are used to form a tree structure called dendrogram representing the hierarchy. The order of the dendrogram is used to order the axis of the heatmaps.

Figure. 5.16 represent the clustering algorithm used in a matrix format on a toy dataset formed for the purpose of this illustration. Figure. 5.17 represents the dendrogram formed for this toy dataset.



	agitation	**anxiety**	**automatism**	**eating**	**energy**
agitation	0	1	5	10	10
anxiety	1	0	5	10	10
automatism	5	5	0	10	10
eating	10	10	10	0	5
energy	10	10	10	5	0

	(agitation, anxiety)	**automatism**	**eating**	**energy**
(agitation, anxiety)	0	5	10	10
automatism	5	0	10	10
eating	10	10	0	5
energy	10	10	5	0

	((agitation, anxiety),automatism)	**(eating, energy)**
((agitation, anxiety),automatism)	0	10
(eating, energy)	10	0

Figure 5.16: Matrix representation of cluster formation on a toy dataset

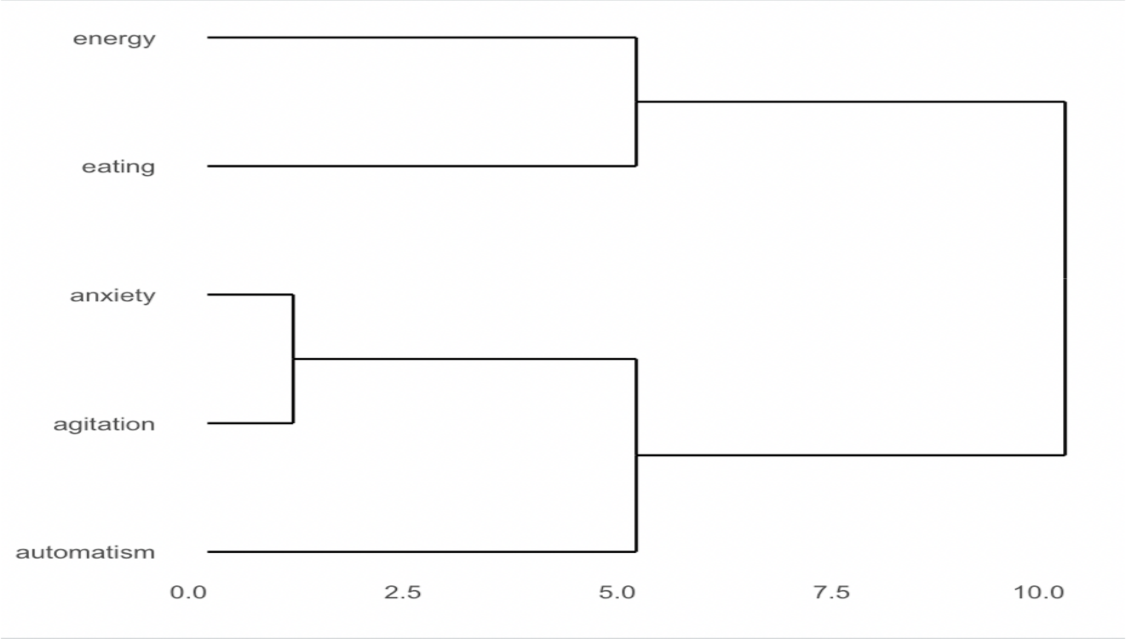


Figure 5.17: Dendrogram formed for the toy dataset

Figure 5.16 represents the distance matrix for the toy dataset and initially every symptom is an individual cluster. The minimum distance is taken to cluster symptoms and after forming a cluster the distance matrix is updated by taking the maximum distance between each element of the previous symptoms clustered. The above process is repeated until all symptoms belong to a cluster forming a dendrogram. The same logic has been used to form the dendrogram for the dataset at hand. From the dendrogram in Figure. 5.17 the clusters formed are energy, eating and anxiety, agitation, automatism. Clusters identified on the dendrogram can be identified on the heatmap with the help of squares on the diagonal.

The heatmaps help in understanding the relation between symptoms along with identifying groups, patterns and highlights the importance of symptom order.

5.6 Visualization of multiple attributes

This section covers visualization of multiple symptoms to identify patterns between them which corresponds to task 4(a) in Chapter 4.

To find patterns between symptoms, one possibility is to use balloon plots for visualizing pairs of symptoms among the selected multiple symptoms. However, in such a case it gets difficult to identify relations beyond two symptoms as the target user's need to look at multiple plots. As, the goal is to visualize all the symptoms at a time, a parallel coordinate is used [22].

Parallel coordinate plot as shown in Figure. 5.18 is generated by selecting symptoms on the stacked bar plot as shown in Figure. 5.7. It helps to simultaneously visualize all selected symptoms. Every symptom is represented by a vertical line and the values on each line represent the severity levels. Although the levels present in the ordinal attribute range from 0 to 4, it is of interest to include the symptoms not reported by a patient and hence -1 is used to represent the same.

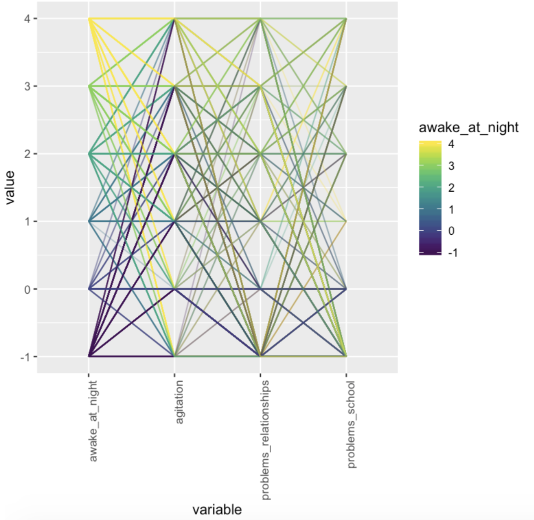


Figure 5.18: Parallel coordinate plot

As can be seen from Figure. 5.18, color has been set to the viridis color palette. Each level from -1 to 4 has a separate color with respect to the first symptom selected. Although we are dealing with ordinal attribute, R considers it to be ordinal numeric due to which a continuous color scale is used which does not affect the results. Using colors to distinguish the different levels with respect to the starting point helps in differentiating them quickly when compared to using a single color. Since multiple patients are involved in a line segment, transparency is used to distinguish the line segments having fewer number of patients and those with more patients.

The target users can interact with the parallel coordinate in two ways, via hovering and brushing which helps in reducing clutter. In R it is not possible to interact and show the changes in the same plot and hence a view, is further split, as shown in Figure. 5.19. The plot on the left is where the target user will interact with the parallel coordinate and the result of the interaction is depicted on the right hand plot due to which color differentiation is not used.

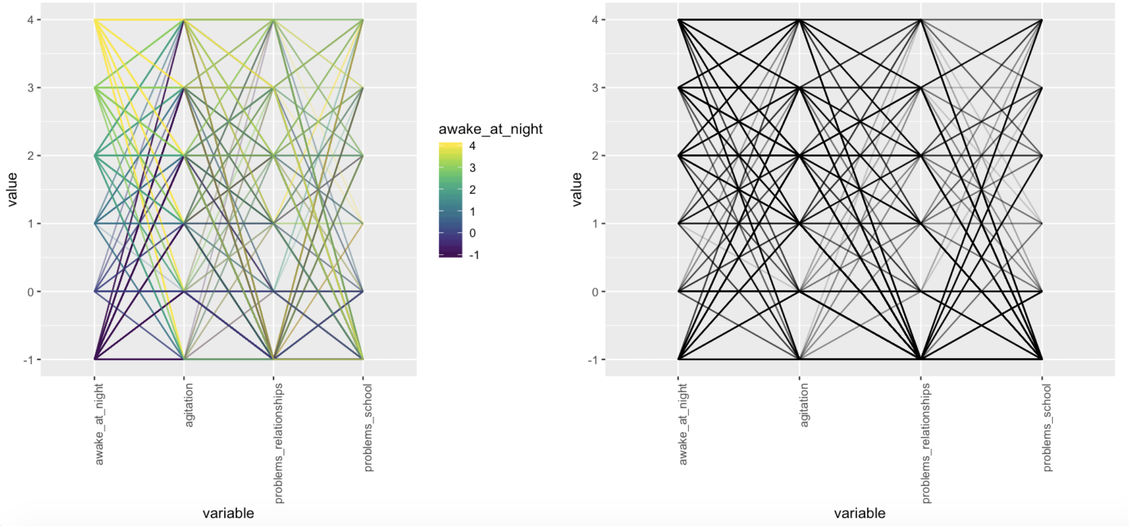


Figure 5.19: Parallel coordinate plot - left plot for interaction and the right plot to reflect the result of interaction

When the target user hovers on the y axis, the lines passing through that symptom and that particular severity level is highlighted as shown in Figure. 5.20. Due to the limitations of R, the target user has to hover exactly over the corresponding symptom and severity level in order to highlight the lines passing through the desired point.

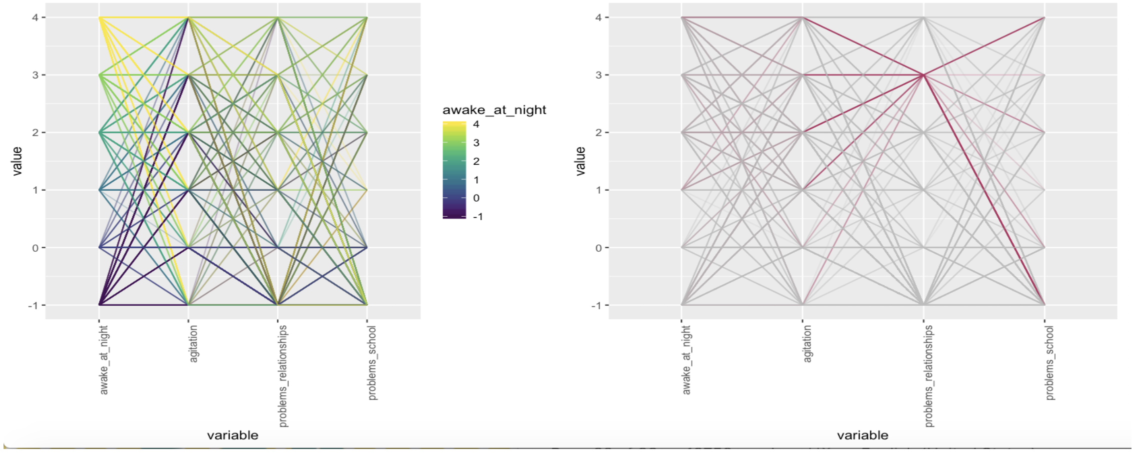


Figure 5.20: Parallel coordinate plots when hovered on the plot on the left

The target user can brush on the parallel coordinate in order to focus on lines passing through the brushed severity levels of a symptom. Figure. 5.21 shows an example of brushing in the parallel coordinate plot. A limitation associated with brushing is that at a time the user can brush over only one symptom and when they brush the next time, results are shown according to the new symptom and severity levels brushed overwriting the previous selection.

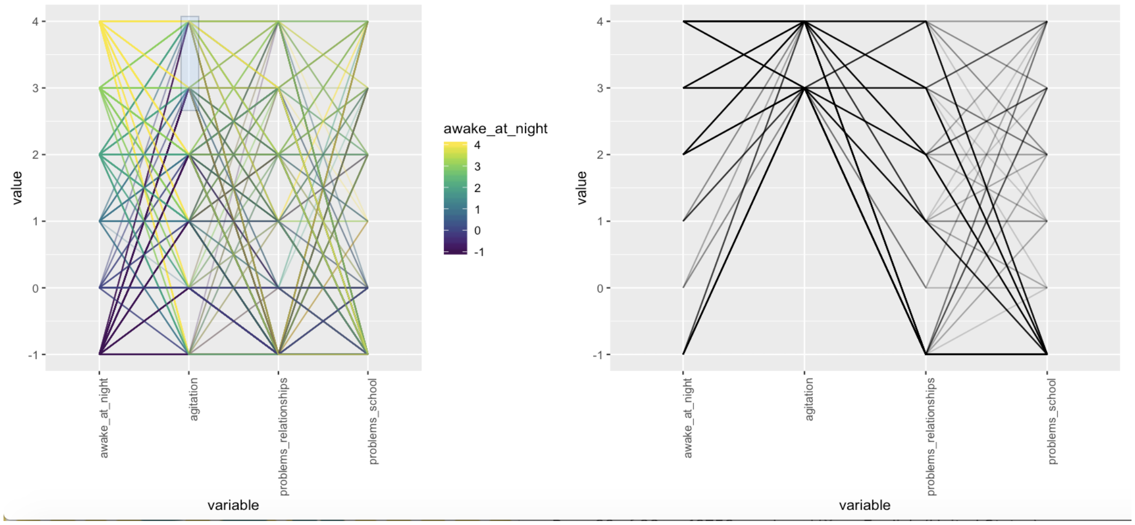


Figure 5.21: Parallel coordinate plots when brushed on the plot on the left

Figure. 5.22 shows the possibility to highlight the desired brushed values.

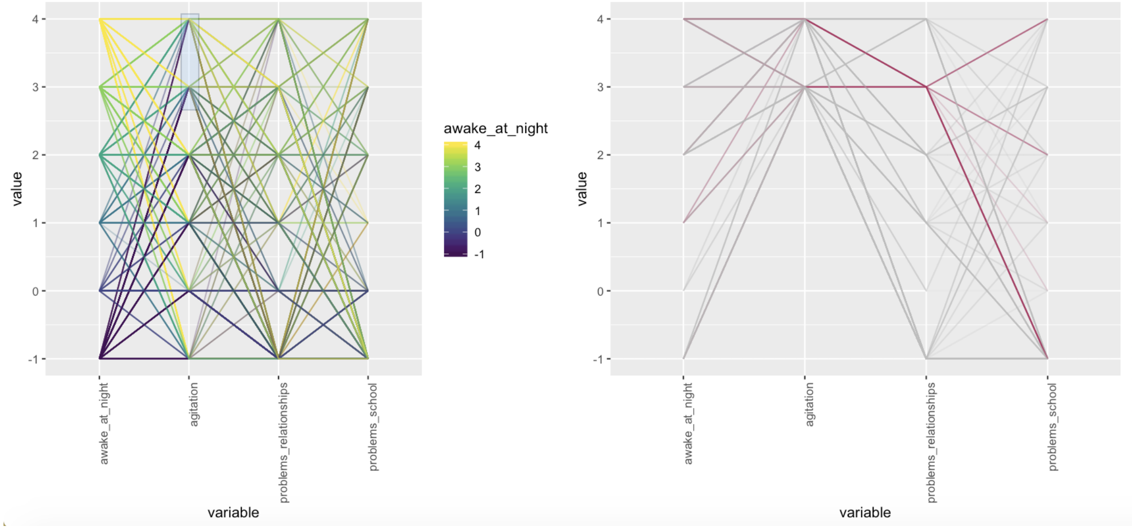


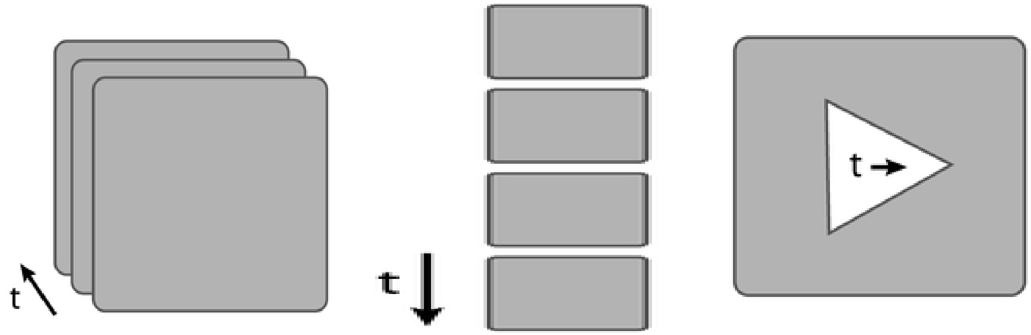
Figure 5.22: Parallel coordinate plots when brushed and hovered on the plot on the left

Parallel coordinate also lets the target user understand how the order in which the symptoms are selected affect the results.

5.7 Comparison of population subsets

This section corresponds to task 5 in Chapter 4 where the task involved is to be able to compare the results obtained through different population subsets with ease. In order to achieve the same, we need to facet the data, meaning we need to split the display. It can either be done by splitting into multiple views or into layers. There are three approaches to faceting and they are superimposition, juxtaposition and explicit encoding as shown in

Figure. 5.23 [25][26].



(a) Superimposition (b) Juxtaposition (c) Explicit encoding

Figure 5.23: Three methods of faceting. (Source: Pep up Your Time Machine – Recommendations for the Design of Information Visualizations of Time-Dependent Data)

Superimposition refers to layering the visualization one on top of the other, whereas juxtaposition refers to displaying the visualizations side by side. Explicit encoding refers to directly visualizing the difference between time points, for example using animation [25].

Superimposition is not a suitable option as layering the visualizations for different population subsets will not help distinguish the differences and similarities easily. Moreover, as the number of layers increases it causes visual clutter. Similarly, explicit encoding does not help as changing views, make it difficult for the users to compare the population subsets and requires the user to recall from memory. Another drawback, is change blindness, wherein users focus on one change and becomes oblivious to the other changes leading to loss of information [22][25].

For the visualization tasks at hand, among the three options, juxtaposition is preferred as it helps in direct comparison letting the users see the differences and similarities between the population subsets clearly. However, it comes with a few drawbacks, like depending on the display area for multiple views and difficulty in comparing visualizations which are far apart. Despite these drawback juxtaposition is more suitable for the project due to the reason mentioned above.

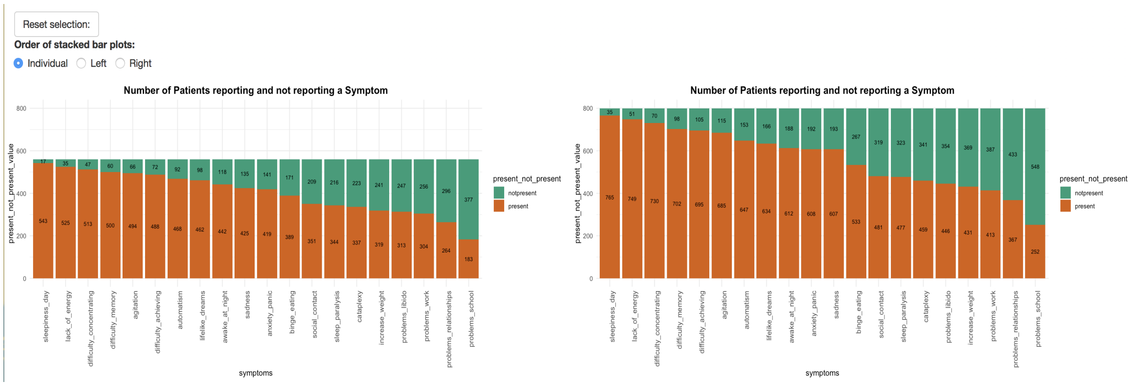
In this project, at a time two different population subsets can be compared due to the display space available. If more than two views are included it would get cluttered making the visualization incomprehensible. According to Tamara Munzner, we are using small multiples as the views for the two population subsets share a common visual encoding while the dataset used are different.

There are comparisons being carried out at the levels of individual, pairs and multiple symptoms for the two population subsets which are explained in subsections 5.7.1, 5.7.2 and 5.7.3.

5.7.1 Comparison of individual symptoms

Distribution of symptoms present and not present

Although the stacked bar plots are ordered in descending order of the lower stacks frequency, when it comes to comparing the two stacked plots it is difficult to map a bar on the left to its corresponding bar on the right and vice-versa. To enable better comparison, the target user is given three options namely individual, left and right to order the stacked bar chart. Figure. 5.24 represents the stacked bar plots when the order selected is individual, it can be seen that the plots retain their original order of being in descending order based on the lower stack frequency.



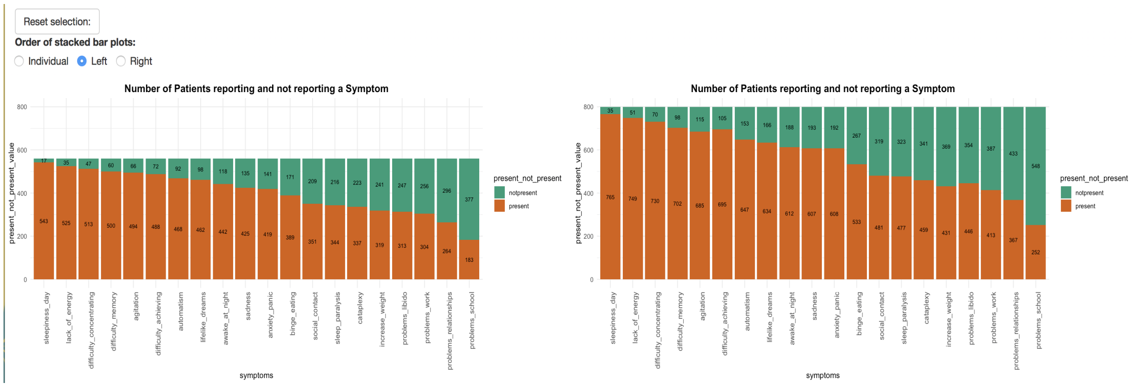
Reset selection:

Order of stacked bar plots:

● Individual ○ Left ○ Right

Figure 5.24: Stacked bar plots ordered based on their individual ordering

The other two options available are left and right, option left implies that the left stacked bars order is used to order both the stacked bar plots as shown in Figure. 5.25 and the right option implies that both the stacked bar plots are ordered using the order of the right stacked bar plot Figure. 5.26.

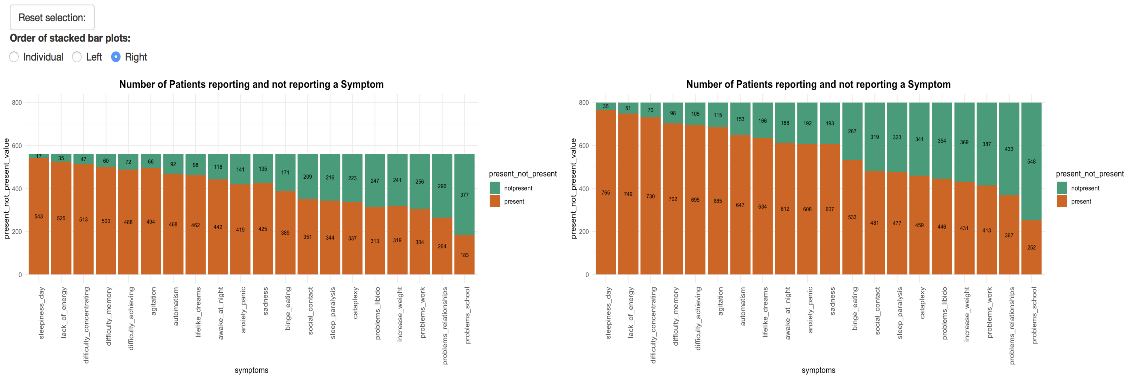


Reset selection:

Order of stacked bar plots:

○ Individual ● Left ○ Right

Figure 5.25: Stacked bar plots ordered based on the left hand side plot's order



Reset selection:

Order of stacked bar plots:

○ Individual ○ Left ● Right

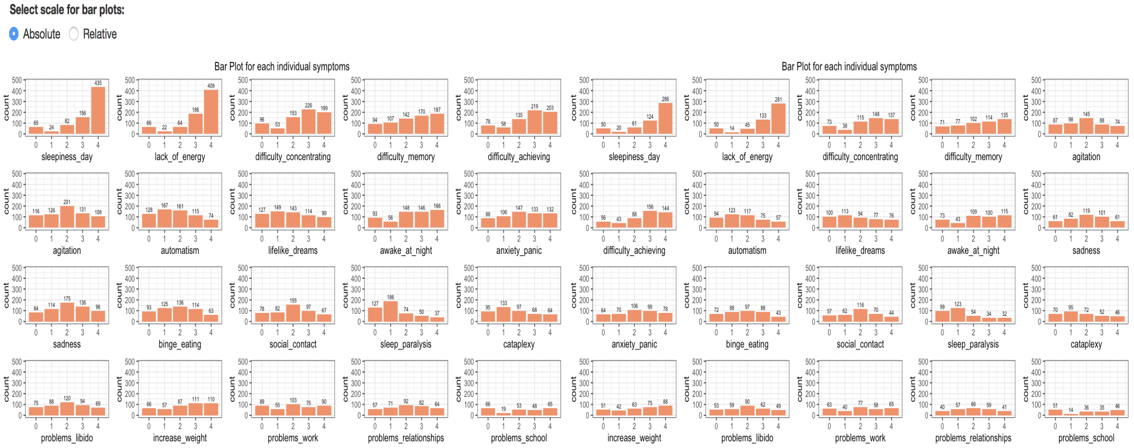
Figure 5.26: Stacked bar plots ordered based on the right hand side plot's order

As can be seen from Figures. 5.24 5.25 5.26, in order to help make better comparison the scales are kept constant, in both the stacked bar plots. The maximum of the total for a bar is taken from either side and is used as the maximum value for the y axis.

Distribution of symptom severity

Figure. 5.27 represents the distribution of symptoms based on severity for two population subsets selected. In order to enable better comparison of the individual bar plots on either side, two scale options are given and they are absolute and relative.

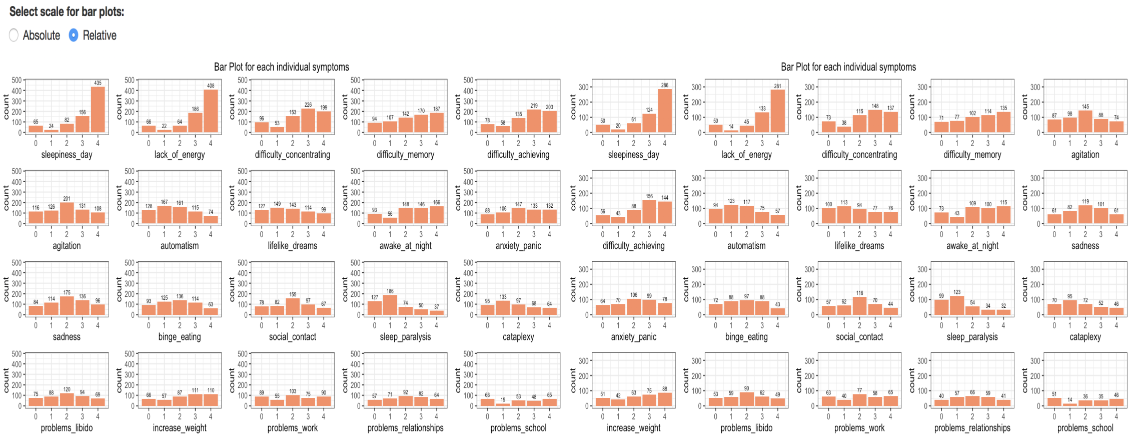
When absolute scale is chosen the max height of a bar among all the plots for the two population subsets is taken as the maximum y axis value as shown in Figure. 5.27. Whereas, when relative is chosen, the maximum height of a bar among the plots on their corresponding population subset is taken and set as the maximum y axis value for the respective population subsets as shown in Figure. 5.28.



Select scale for bar plots:

● Absolute ○ Relative

Figure 5.27: Individual bar plots depicting the distribution of symptoms in terms of severity with absolute scale

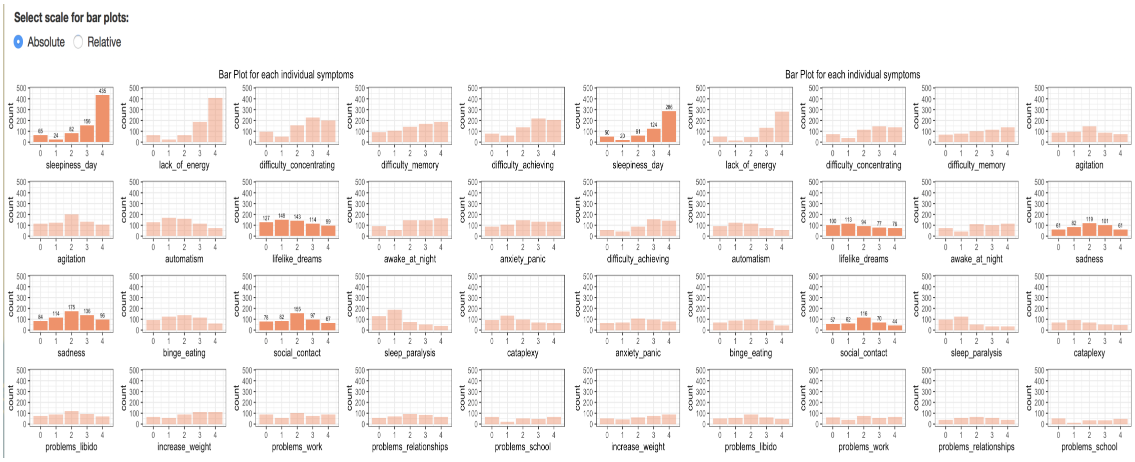


Select scale for bar plots:

○ Absolute ● Relative

Figure 5.28: Individual bar plots depicting the distribution of symptoms in terms of severity with relative scale

When a symptom is clicked on the stacked bar plot, the corresponding symptom is highlighted on both the population subsets as shown in Figure. 5.29. The highlighting is achieved by reducing the saturation of plots corresponding to symptoms not selected.



Select scale for bar plots:

● Absolute ○ Relative

Figure 5.29: Individual bar plots depicting the distribution of symptoms with symptoms highlighted corresponding to the symptoms selected in the stacked bar plot.

5.7.2 Comparison of pairs of symptoms

It would be useful if the target users can compare the results for the two population subsets with ease. Hence, similar to the stacked bar plots, the user can rearrange the x and y axis of the heatmaps based on individual, agreement-left, association-left, agreement-right and association-right. The following paragraph explains each of these options for two different population subsets when the order of heatmap is set to value.

Since, the range can vary for different subsets of the dataset, the agreement is reported in terms of percentage where agreement between every pair of symptom is divided by the maximum agreement. This step aids in the comparison process between two different subsets of the data.

Individual refers to the original ordering of the heatmaps and Figure. 5.30 represents the same.

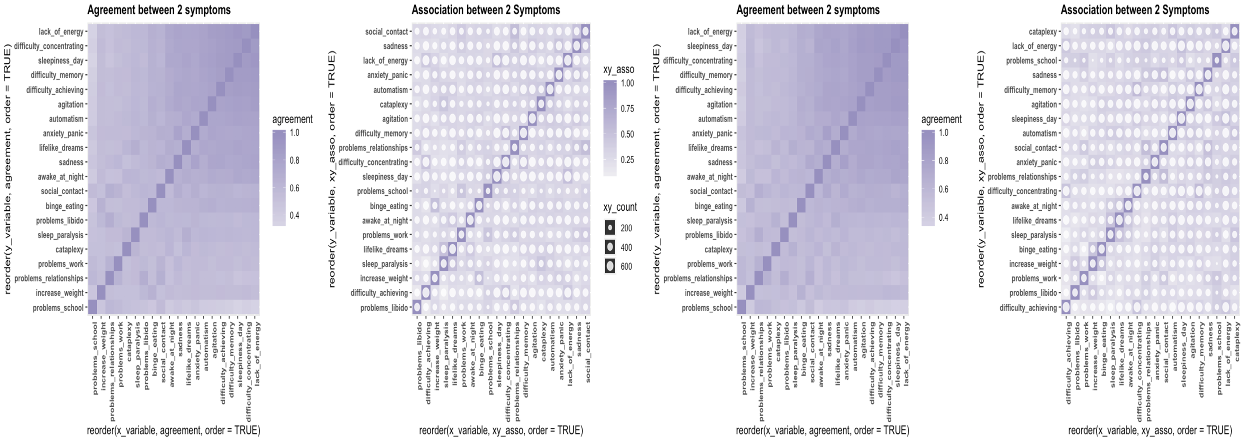


Figure 5.30: Heatmaps reordered using option - Individual

Agreement-left, takes the order of the agreement heatmap on the left and reorders the other heatmaps accordingly, Figure. 5.31 gives an example of agreement-left.

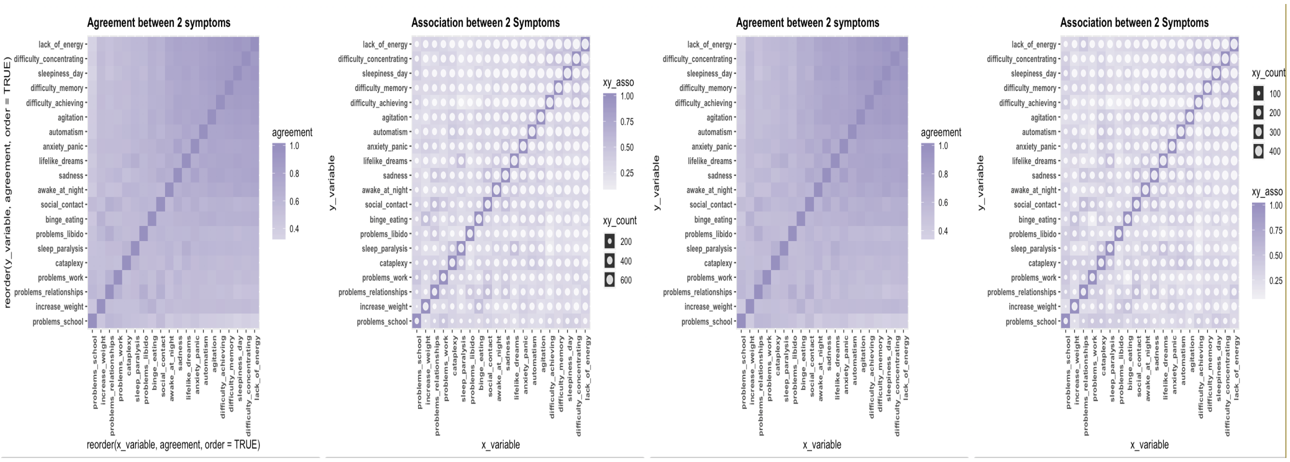


Figure 5.31: Heatmaps reordered using Agreement-left

Association-left, takes the order of the association plot on the left hand side and uses it to reorder all other heatmaps as shown in Figure. 5.32.

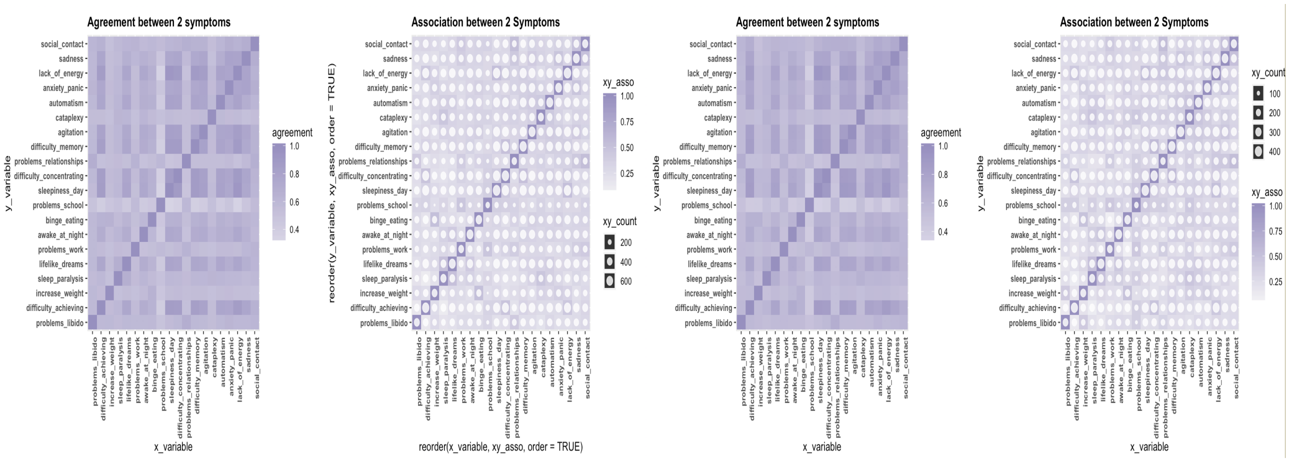


Figure 5.32: Heatmaps reordered using Association-left

.

Agreement-right, uses the order of the agreement plot on the right hand side and reorders the rest of the heatmaps accordingly as shown in Figure. 5.33.

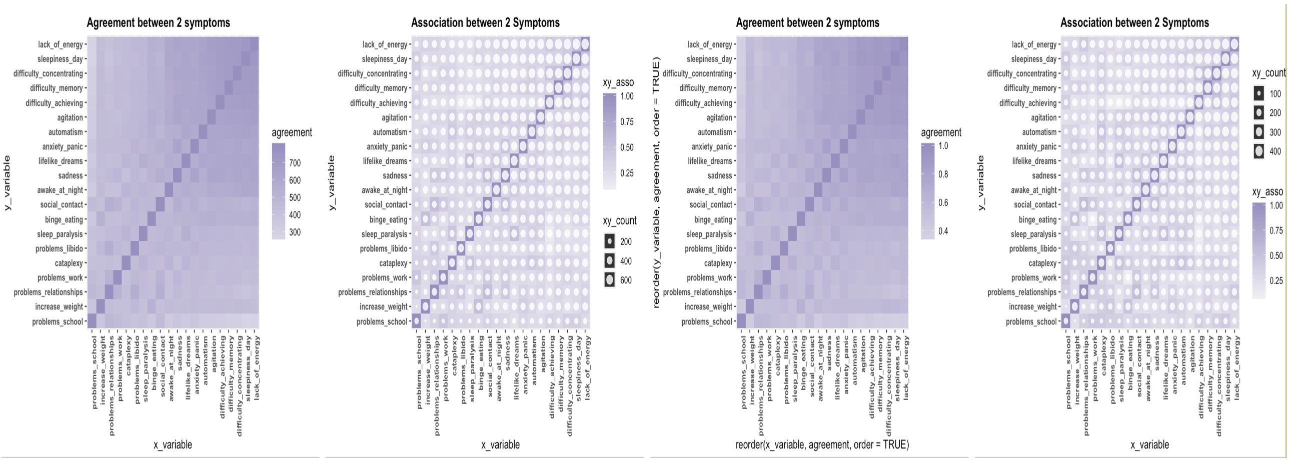


Figure 5.33: Heatmaps reordered using Agreement-right

Lastly, association-right, takes the order from the association heatmap on the right and reorders the other heatmaps accordingly as shown in Figure. 5.34.

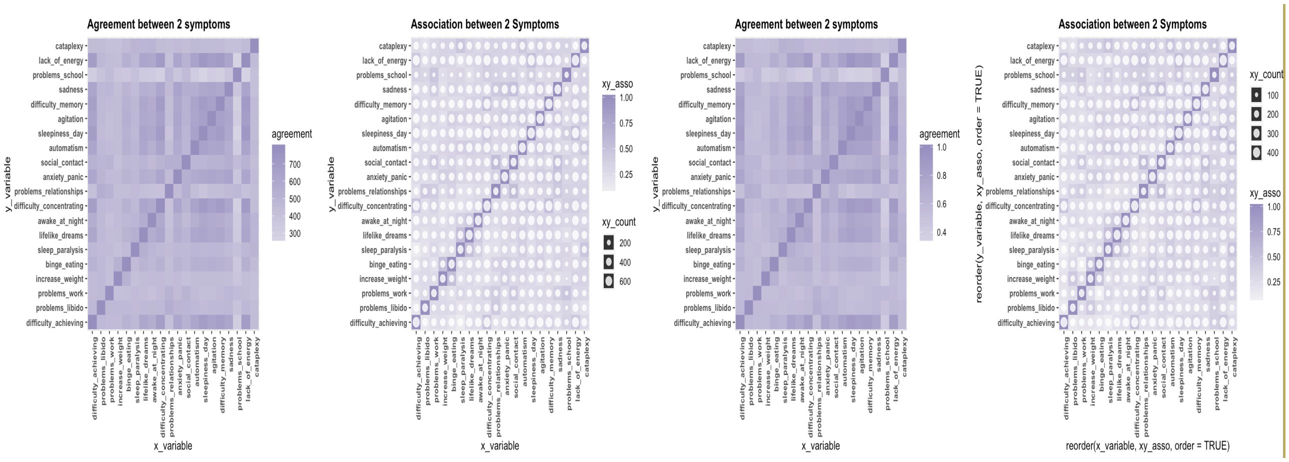
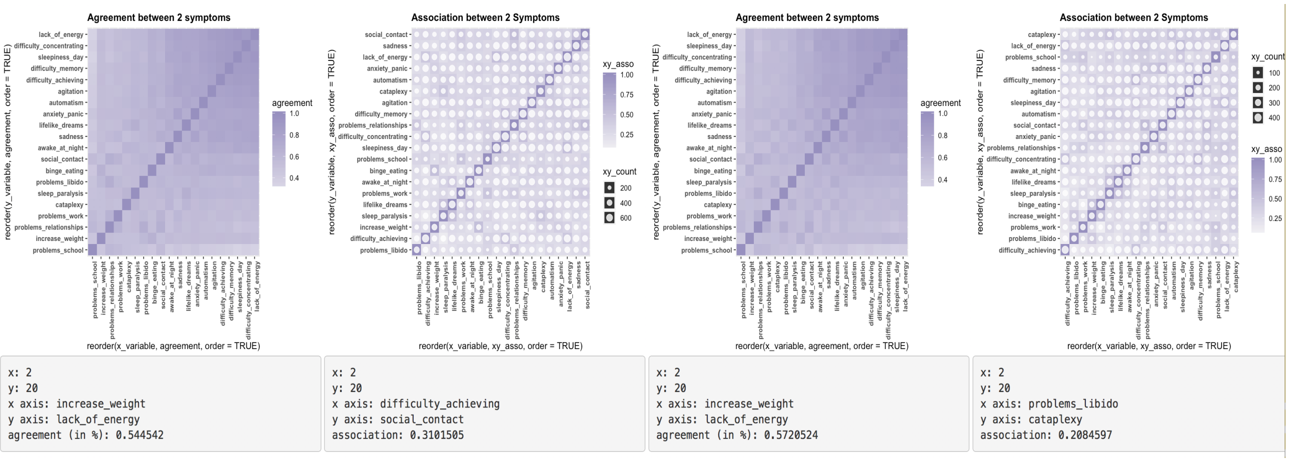


Figure 5.34: Heatmaps reordered using Association-right

The interactions possible with the heatmap include hovering and clicking. When the target user hovers over any tile on the heatmap, the corresponding symptoms should be ideally highlighted on the axis. As it is not possible in R to access a single axis label, the corresponding information is provided in a text box. For the other heatmaps the corresponding information pertaining to the respective heatmap is printed as shown in Figure. 5.35.


Agreement between 2 symptoms
Association between 2 Symptoms
Agreement between 2 symptoms
Association between 2 Symptoms
agreement
1.0
0.8
0.6
0.4
xy_asso
1.00
0.75
0.50
0.25
xy_count
200
400
600
xy_count
100
200
300
400
reorder(y_variable, agreement, order = TRUE)
reorder(y_variable, xy_asso, order = TRUE)
reorder(x_variable, agreement, order = TRUE)
reorder(x_variable, xy_asso, order = TRUE)

```
x: 2
y: 20
x axis: increase_weight
y axis: lack_of_energy
agreement (in %): 0.544542
```

```
x: 2
y: 20
x axis: difficulty_achieving
y axis: social_contact
association: 0.3101505
```

```
x: 2
y: 20
x axis: increase_weight
y axis: lack_of_energy
agreement (in %): 0.5720524
```

```
x: 2
y: 20
x axis: problems_libido
y axis: cataplexy
association: 0.2084597
```

Figure 5.35: Result when hovered over the heatmaps

The concept of detail on demand is used when the target user clicks on the heatmap [22]. When clicked the balloon plots are displayed for the two symptoms selected. Unlike hovering, in this case all the plots are displayed for the symptoms that is being clicked, as the goal is to enable comparison. Figure. 5.36 depicts the balloon plots for two different population subsets for the symptoms increase in weight and problems school.

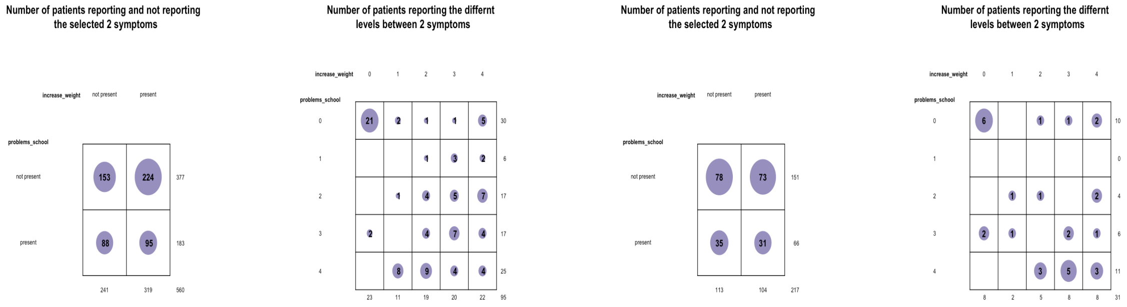


Figure 5.36: Balloon plots pertaining to two different subsets

5.7.3 Comparison of multiple symptoms

Figure. 5.37 depicts the parallel coordinate plots for two population subsets selected. Although the interactions, brushing and highlighting is done on the plot corresponding to one population subset, the results are reflected on the plots for both the population subsets. Similarly, when selecting symptoms for the parallel coordinate plot on the stacked bar plot, the same symptoms are used for both the population subsets to enable better comparison.

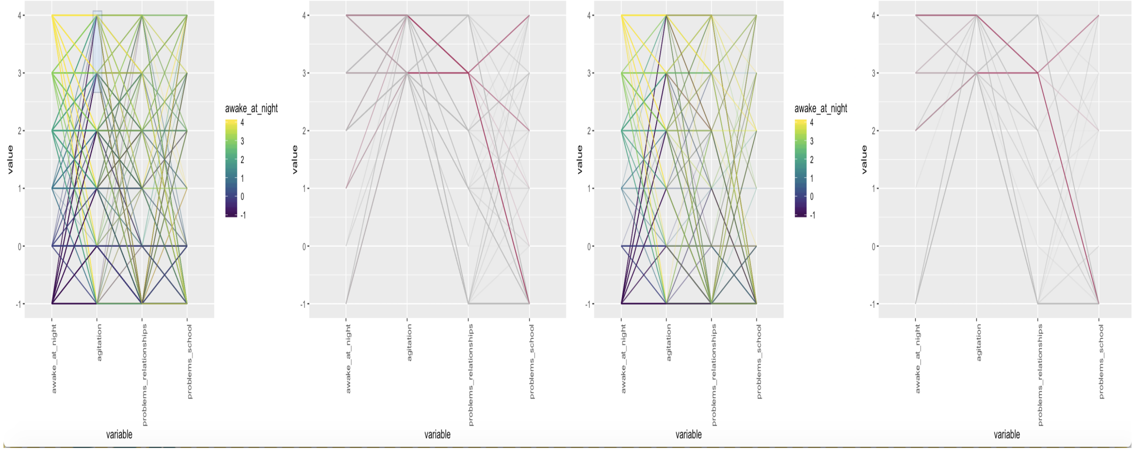
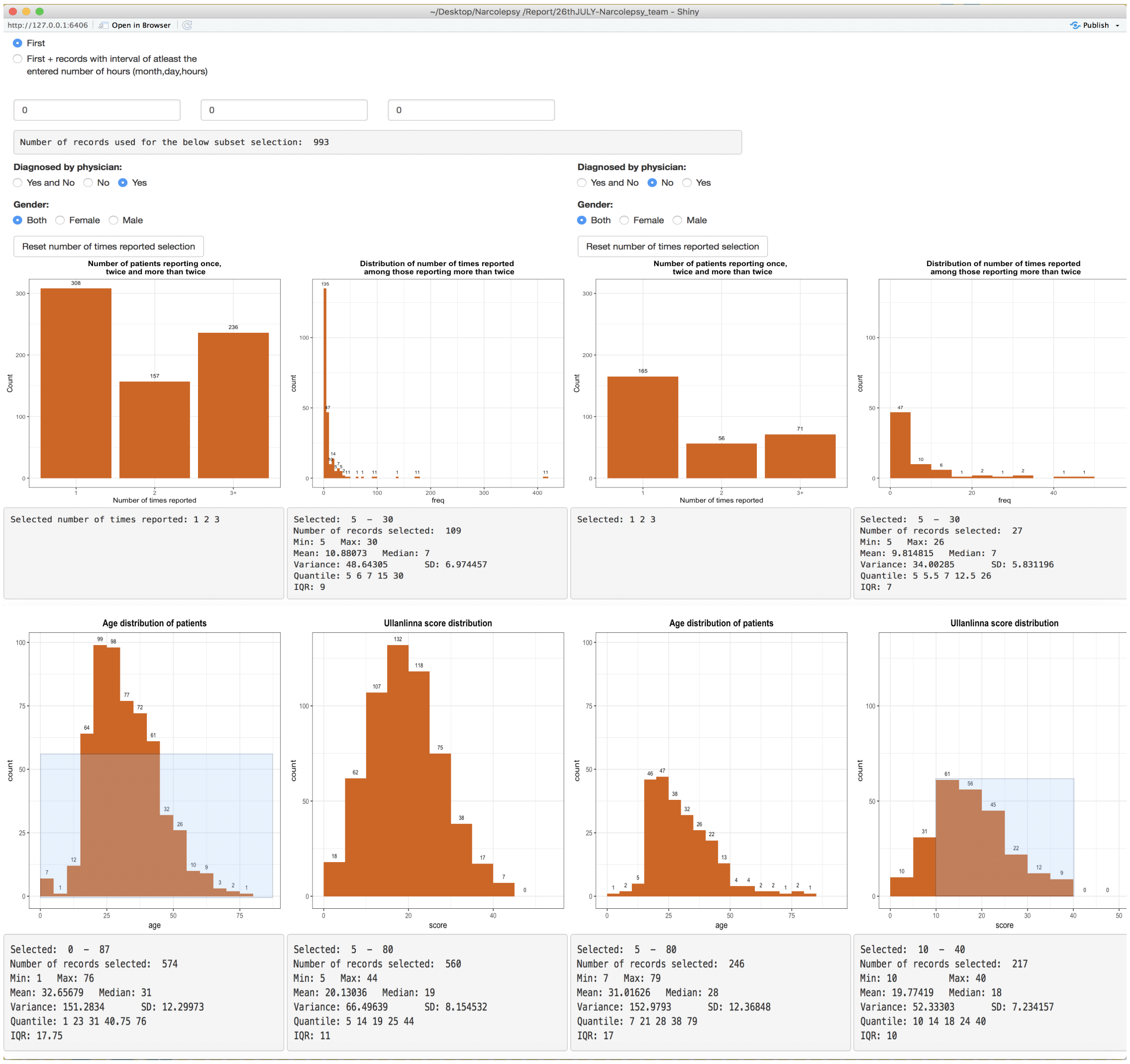


Figure 5.37: Parallel coordinate plots for the two different subsets selected

5.8 Dashboard

This section depicts the dashboards developed. For a summary of the dashboard details please refer to Appendix A.

Figure. 5.38 depicts the visualization developed for filtering records and selecting population subsets.



~/Desktop/Narcolepsy /Report/26thJULY-Narcolepsy_team - Shiny

http://127.0.0.1:6406 | Open in Browser | Publish

● First
○ First + records with interval of atleast the entered number of hours (month,day,hours)

0 | 0 | 0

```
Number of records used for the below subset selection:  993
```

Diagnosed by physician:
○ Yes and No ○ No ● Yes

Gender:
● Both ○ Female ○ Male

Reset number of times reported selection

Diagnosed by physician:
○ Yes and No ● No ○ Yes

Gender:
● Both ○ Female ○ Male

Reset number of times reported selection

```
Selected number of times reported: 1 2 3
```

```
Selected:  5  -  30
Number of records selected:  109
Min: 5   Max: 30
Mean: 10.88073   Median: 7
Variance: 48.64305       SD: 6.974457
Quantile: 5 6 7 15 30
IQR: 9
```

```
Selected: 1 2 3
```

```
Selected:  5  -  30
Number of records selected:  27
Min: 5   Max: 26
Mean: 9.814815   Median: 7
Variance: 34.00285       SD: 5.831196
Quantile: 5 5.5 7 12.5 26
IQR: 7
```

```
Selected:  0  -  87
Number of records selected:  574
Min: 1   Max: 76
Mean: 32.65679   Median: 31
Variance: 151.2834       SD: 12.29973
Quantile: 1 23 31 40.75 76
IQR: 17.75
```

```
Selected:  5  -  80
Number of records selected:  560
Min: 5   Max: 44
Mean: 20.13036   Median: 19
Variance: 66.49639       SD: 8.154532
Quantile: 5 14 19 25 44
IQR: 11
```

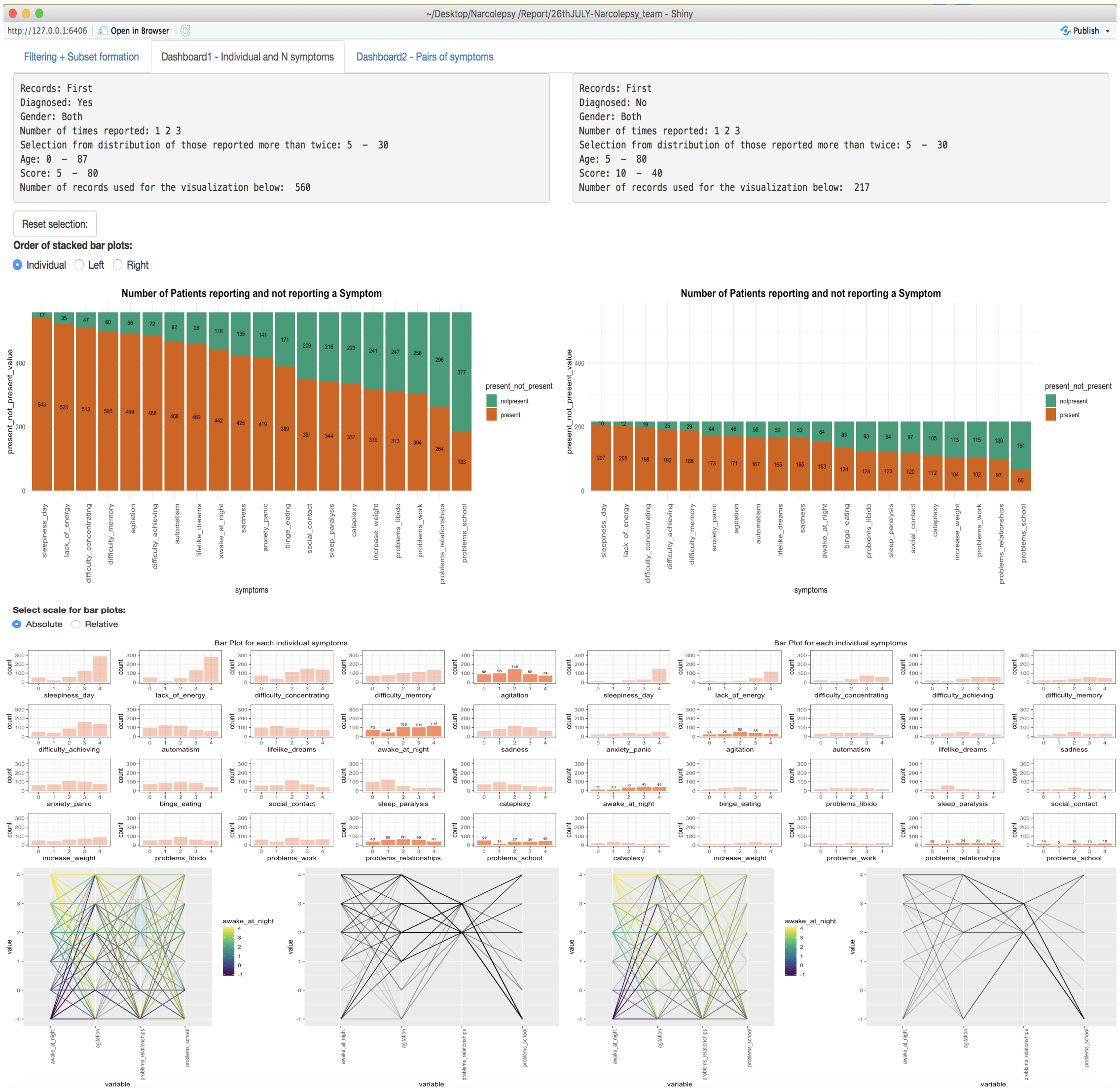
```
Selected:  5  -  80
Number of records selected:  246
Min: 7   Max: 79
Mean: 31.01626   Median: 28
Variance: 152.9793       SD: 12.36848
Quantile: 7 21 28 38 79
IQR: 17
```

```
Selected:  10  -  40
Number of records selected:  217
Min: 10          Max: 40
Mean: 19.77419   Median: 18
Variance: 52.33303       SD: 7.234157
Quantile: 10 14 18 24 40
IQR: 10
```

Number of patients reporting once, twice and more than twice
Count
Number of times reported
1
2
3+
308
157
236
Distribution of number of times reported among those reporting more than twice
count
freq
135
47
Number of patients reporting once, twice and more than twice
Count
Number of times reported
165
56
71
Distribution of number of times reported among those reporting more than twice
count
freq
47
10
6
Age distribution of patients
count
age
99
98
77
72
64
61
32
26
12
7
Ullanlinna score distribution
count
score
132
118
107
75
62
38
18
17
7
0
Age distribution of patients
count
age
47
46
38
32
26
22
13
Ullanlinna score distribution
count
score
61
56
45
31
22
12
10
9

Figure 5.38: Filtering and Subset selection

Figure. 5.39 depicts the visualization developed for individual and multiple symptoms.

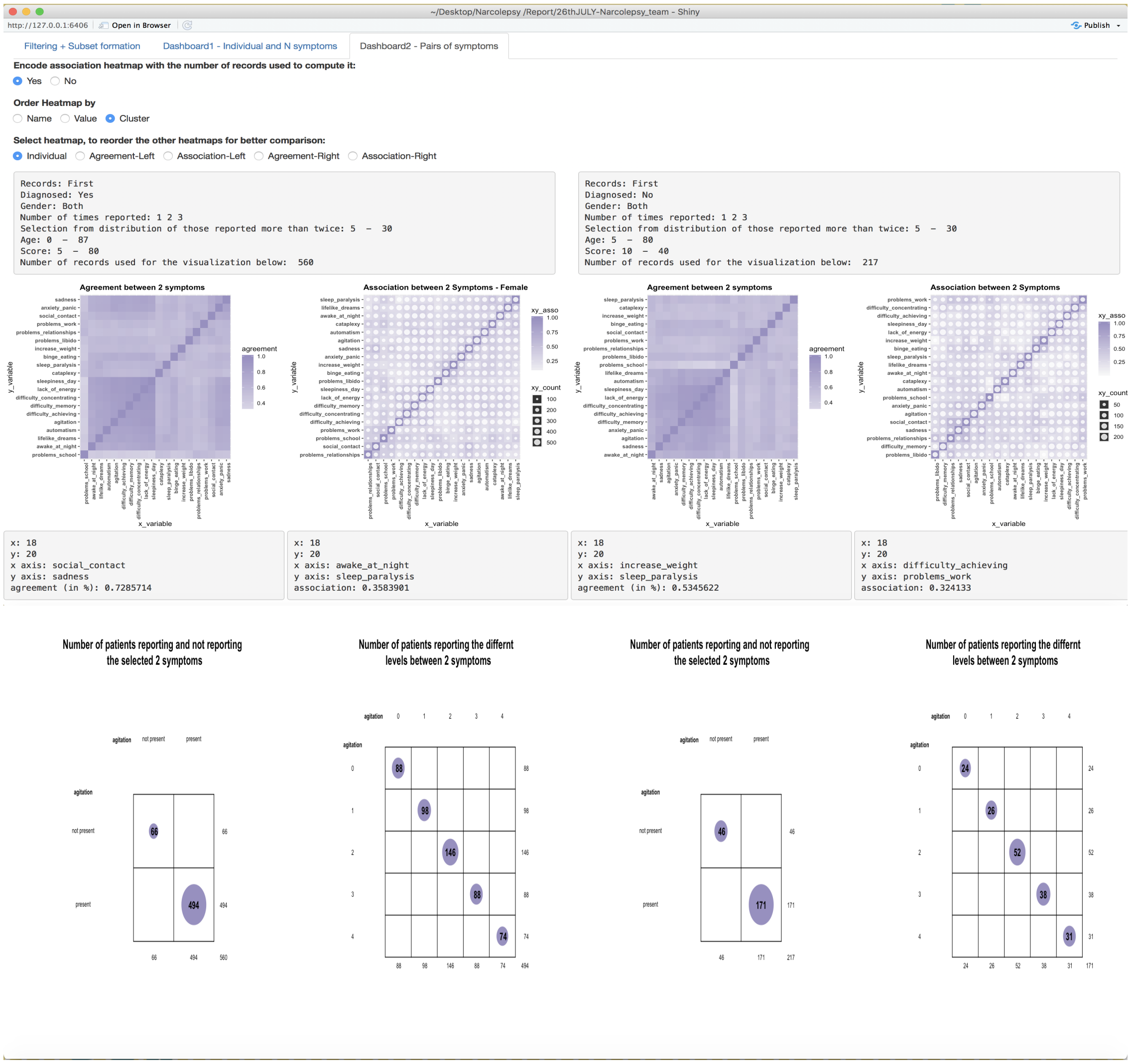

~/Desktop/Narcolepsy /Report/26thJULY-Narcolepsy_team - Shiny
http://127.0.0.1:6406
Open in Browser
Publish
Filtering + Subset formation
Dashboard1 - Individual and N symptoms
Dashboard2 - Pairs of symptoms

```
Records: First
Diagnosed: Yes
Gender: Both
Number of times reported: 1 2 3
Selection from distribution of those reported more than twice: 5  -  30
Age: 0  -  87
Score: 5  -  80
Number of records used for the visualization below:  560
```

```
Records: First
Diagnosed: No
Gender: Both
Number of times reported: 1 2 3
Selection from distribution of those reported more than twice: 5  -  30
Age: 5  -  80
Score: 10  -  40
Number of records used for the visualization below:  217
```

Figure 5.39: Individual and multiple symptoms

Figure. 5.40 represents the visualization developed for pairs of symptoms.


~/Desktop/Narcolepsy /Report/26thJULY-Narcolepsy_team - Shiny
http://127.0.0.1:6406
Open in Browser
Publish
Filtering + Subset formation
Dashboard1 - Individual and N symptoms
Dashboard2 - Pairs of symptoms

Encode association heatmap with the number of records used to compute it:
● Yes ○ No

Order Heatmap by
○ Name ○ Value ● Cluster

Select heatmap, to reorder the other heatmaps for better comparison:
● Individual ○ Agreement-Left ○ Association-Left ○ Agreement-Right ○ Association-Right

```
Records: First
Diagnosed: Yes
Gender: Both
Number of times reported: 1 2 3
Selection from distribution of those reported more than twice: 5  –  30
Age: 0  –  87
Score: 5  –  80
Number of records used for the visualization below:  560
```

```
Records: First
Diagnosed: No
Gender: Both
Number of times reported: 1 2 3
Selection from distribution of those reported more than twice: 5  –  30
Age: 5  –  80
Score: 10  –  40
Number of records used for the visualization below:  217
```

Agreement between 2 symptoms
Association between 2 Symptoms - Female
Agreement between 2 symptoms
Association between 2 Symptoms
y_variable
x_variable
agreement
1.0
0.8
0.6
0.4
xy_asso
1.00
0.75
0.50
0.25
xy_count
100
200
300
400
500
50
150

```
x: 18
y: 20
x axis: social_contact
y axis: sadness
agreement (in %): 0.7285714
```

```
x: 18
y: 20
x axis: awake_at_night
y axis: sleep_paralysis
association: 0.3583901
```

```
x: 18
y: 20
x axis: increase_weight
y axis: sleep_paralysis
agreement (in %): 0.5345622
```

```
x: 18
y: 20
x axis: difficulty_achieving
y axis: problems_work
association: 0.324133
```

Number of patients reporting and not reporting the selected 2 symptoms
Number of patients reporting the differnt levels between 2 symptoms
Number of patients reporting and not reporting the selected 2 symptoms
Number of patients reporting the differnt levels between 2 symptoms
agitation
not present
present
66
494
560
0
1
2
3
4
88
98
146
74
46
171
217
24
26
52
38
31

Figure 5.40: Pairs of symptoms

5.9 Implementation

In this section we cover the algorithm level, primarily focusing on efficiently handling the visual encoding and interactions. The algorithm level concerns the computational issues when compared to the visual encoding and interaction level, where the concern is with regard to human perception [22]. In this level the focus is on the correctness of the algorithm and its speed.

The choice of visual encoding and interaction discussed in detail in this chapter is evaluated and the results of the evaluation in discussed in Chapter 6. Therefore, this section discusses the performance of the visualization in terms of speed.

When the target user interacts with the visualization, the primary goal is to ensure that the interactions are smooth and do not take a lot of time to reflect. However, in this case it takes a few seconds when compared to taking a few milliseconds. The primary reason behind the delay is due to the choice of implementation, which is R in this case. In R, when we feed in the data, it goes through all the records, picks the relevant records for a particular task and then generates the visualization and hence the delay.

In this project ggplot2 [11] package is used for the plots and using shiny [12] package, interactions are made possible on the plots. However, when alternatives to ggplot2 like plotly are considered they provide interactions only with respect to viewing information on the plot and do not work well with shiny to make selections such that other plots can be modified. Therefore, when an interaction is performed, R first needs to understand the location, find the symptoms and values involved to make the required changes which is done programmatically. The associated actions with regard to an interaction are also programmed and not automated, causing delays.

[11] https://ggplot2.tidyverse.org/

[12] https://shiny.rstudio.com/

Chapter 6

Evaluation

In order to evaluate the visual analytics framework discussed in Chapter 5, a user study was conducted and this chapter explains the design of the study, the participants involved and the results obtained.

6.1 Method and participants

The user study was conducted with three participants from the sleep medicine center, Kempenhaeghe who's brief profile is as follows,

1. User 1 - somnologist, coordinator scientific research
2. User 2 - child neurologist, somnologist
3. User 3 - psychologist, somnologist and PhD candidate

The visualization was evaluated with the researchers who were lead through all the features and interactions available in the tool to measure, the perceived usefulness and ease of use.

During the user study, the participants were given a questionnaire which consisted of questions in the form of statements. The responses were recorded using a likert scale ranging from 1 to 5, where 1 stands for strongly disagree and 5 stands for strongly agree. It also contained a few open ended questions. The questionnaire aimed to check what features were useful, what was missing and if the tool was suitable for wide adoption.

The questions were designed in google form and was given to the users prior to commencement of the user study. The code was shared with the users on 10th July 2020 with an installation guide. Two of the users had the program installed in their system. The user study was conducted online on 23rd July 2020 which went on for about 2 hours and was recorded. The dataset was well known to the users. The study was conducted simultaneously with all the three participants and one of them was interacting with the visualization and all the queries raised were resolved. While interacting with the visualization, corresponding sections of the questionnaire were filled.

6.2 Results

This section discusses the results obtained through the user study, for the exact questionnaire and response received refer to Appendix B and Appendix C respectively. Table. 6.1 represents the color scheme used to depict the results.

1	2	3	4	5

Table 6.1: Color scheme used to represent the results obtained

6.2.1 Selection mechanism

Table. 6.2 lists the questions and responses received under selection mechanism pertaining to task 1 in Chapter 4.

Question	User 1	User 2	User 3
Required attributes are covered.	2	3	4
Interactions are clear.	4	4	4
The visualization helps in understanding the population better.	5	4	4

Table 6.2: User study questions and responses for selection mechanism

From Table. 6.2 is can be seen that the interactions and distribution plots were clear and served useful in understanding the dataset population better. To an open ended question on what attributes they found missing in the selection mechanism, users 1 and 2 stated they missed selection of records based on symptoms and hence the reason for a low score.

6.2.2 Visualizing individual symptoms

Table. 6.3 represents the questions and the response received under visualizing individual symptoms pertaining to tasks 2(a) and 2(b) from Chapter 4.

Question	User 1	User 2	User 3
Stacked bar plot helps in seeing the distribution clearly.	5	5	5
Visualizing all individual symptoms at a time is useful.	4	4	5
Identifying trends in severity distribution is straightforward.	4	4	4
Ordering individual bar plots according to stacked bar is useful.	5	4	5
Distribution plots aid in selection of multiple symptoms.	4	4	5

Table 6.3: User study questions and responses for visualizing individual symptoms

From Table. 6.3 it can be concluded that the plots help to see the distribution clearly and infer new knowledge which helps in taking further decisions. The choice to align the stacked bar plot and the individual bar plots were found beneficial. When an ended question was asked as to how the distribution plots help in the selection of multiple symptoms, they responded saying *"The exploritary function is very helpfulll. The insight definetely* *helps to make further decisions."*, *"clinically we only have a very rough idea, based on presence or absence of a small number of symptoms. All the visualisation info is new; e.g. the distribution of burden; but also the relevance of certain symptoms."*.

6.2.3 Visualizing pairs of symptoms

Table. 6.4 depicts the questions and responses received under visualizing pairs of symptoms pertaining to tasks 3(a) and 3(b) in Chapter 4.

Question	**User 1**	**User 2**	**User 3**
Heatmaps help in identifying groups of related symptoms.	4	4	4
Heatmaps help in identifying symptoms with low and high association.	4	4	5
Option to encode number of records used for association heatmap is useful.	4	4	4

Table 6.4: User study questions and responses for visualizing pairs of symptoms

From Table. 6.4 it can be concluded that the heatmap helped in identifying groups and trends while the option to encode the association heatmap with the number of records helped in interpreting the reliability of association which was the goal. However, it was suggested that the option to encode the number of records used to compute association could have been given in a simpler form like *"show number of records: yes/no"*.

6.2.4 Visualizing multiple symptoms

Table. 6.4 gives an overview of the questions and responses received under visualizing multiple symptoms pertaining to task 4(a) in Chapter 4.

Question	**User 1**	**User 2**	**User 3**
Visualization helps in identifying patterns.	4	2	2
Highlighting helps in identifying patterns.	2	2	3
Brushing helps in identifying patterns.	2	2	4

Table 6.5: User study questions and responses for visualizing multiple symptoms.

As can be seen from the responses in Table. 6.5 identifying patters when visualizing multiple symptoms at a time has got mixed response. Overall, it can be concluded that the visualization for multiple symptoms needs to be improved to help identify patterns.

The users missed seeing patterns as the number of records per line segment was not available during interaction. This is not possible in R as the width cannot be set for each line segment based on number of records involved. Similarly, when hovering on the line segment, R returns a point and it is not possible to trace the beginning and ending of the line segment to compute the number of records. Another reason, is the familiarity with interpreting parallel coordinates and it could be the case where a learning curve is required.

6.2.5 Comparison

Table. 6.6 represents the questions and response received under comparing the two population subsets with regard to task 5 in Chapter 4.

Question	**User 1**	**User 2**	**User 3**
Identify differences between two population subsets in individual symptoms.	5	4	5
Identify differences between two population subsets in pairs of symptoms.	4	4	4
Helps in comparing groups of symptoms.	4	4	3

Table 6.6: User study questions and responses for comparing two population subsets.

The users were asked an open ended question as to why comparing groups of symptoms was useful or not to which they responded saying, *"we do not know much about interactions of symptoms; clusters etc. So everything is new /useful"*, *"The function of exploration is very useful to gain new hypothesis"*, *"The possibility to see the graphs next to each other is helpful to get an overview over correlations and percentage of symptoms. The associations between symptoms is more difficult to grab in one glance as these are complex data, so also when two subsets are visible."*.

From Table. 6.6 it can be concluded that the visualization helped compare the two population subsets which was the goal.

6.2.6 Overall visualization

In this section, the evaluation comprised of two open ended questions, one what the participants liked about the visualization and the other was what could be improved in the visualization.

In response to what they liked about the visualization they said, they found it to give inspiration, ideas and new insights which helps in forming new hypothesis. Other points include the ability to look at the data in one glance and being able to compare the two subsets easily.

In response to what could be improved, they commented, the overall speed of the program is slow. Further they expressed their desired to have a pre-selection of attributes to eliminate records not of interest. One of the users wished to have an absolute and relative scale in the stacked bar plot. In parallel coordinate plot the lines colored yellow were not clearly visible and needed to be improved. In heatmaps it was found that identifying symptoms with low association was difficult due to the choice of color used.

Chapter 7

Conclusion

This project aimed at building a visual analytics tool to understand the symptoms related to narcolepsy a sleep disorder. This project was done in collaboration with researchers from sleep medicine center, Kempenhaeghe. It was mutually agreed with the researchers the study would focus on the symptom spectrum to understand its influence on narcolepsy.

7.1 Summary

The goal here is to identify and establish relations between symptoms independent of the temporal aspects and each record is treated as an independent item. This project follows the Tamara Munzners 4 nested levels of visualization consisting of domain, task, visual encoding and algorithm.

Primarily the focus of the visualization is to form interesting population subsets. Subsequently, the visualization tasks were further broken down to analyse distribution of individual symptoms, establish relations between pairs of symptoms based on occurrence, look for patterns among multiple symptoms and lastly compare the population subsets to identify similarities and differences.

The visualization was implemented using R programming language and a customized interactive dashboard was built. The key features of this tool include design of different plots, giving suitable options to scale, ordering of plots to identify groups and trends and developing simultaneous views to facilitate comparison.

The visualization tool has been initially evaluated by three researchers through a user study to check if the tool helps in gaining new insights and forming hypothesis. The evaluation covered five sections and the conclusions are briefly described as follows, selection mechanism overall met the user expectation but they wished to see more selection options. Visualization of individual and pair of symptoms were found extremely useful to find new insights. Visualization of multiple symptoms received mixed response and improvements are required to help identify patterns. Lastly, comparison of population subsets was considered useful. It can be concluded that the project was successful in achieving majority of the tasks in the exploration and formation of hypothesis in the study of narcolepsy.

7.2 Future work

This project is the first step towards developing an extensive visual analytics framework for narcolepsy with a specific focus to gain a broader understanding of the symptom spectrum. While this project focused on visualizing the symptom spectrum on the whole, it can be further extended to visualize individual patient records to give personalized treatment. Currently, the project was evaluated by three researchers, however over time as more researchers use the visualization tool, new perspective will evolve for further developments like including pre-selection and improving the visualization for multiple symptoms. The parallel plots can be improved further by adding a feature to indicate number of records in each line segment using width or printing them.

Appendix A

Dashboard

All the visualizations developed for the tasks needs to be in one interactive interface for which a dashboard is developed. The dashboard consists of three tabs, one for selecting the records to be considered and forming the population subsets of choice, second for visualizing individual symptoms along with multiple symptoms and lastly the third tab for visualizing pairs of symptoms. The visualization included in each tab is as follows,

Tab 1 is called Filtering + Subset formation, which lets the target users select the records they want to consider followed by subset selection. Under subset selection the target user can form two interesting population subsets of choice using the attributes, diagnosed by physician, gender, number of times reported, number of time reported when reporting more than twice, age and ullanlinna score.

Therefore, this tab covers the task to form interesting population subsets based on relevant attributes and also helps the target users understand the population better, by letting them compare different groups of the dataset population.

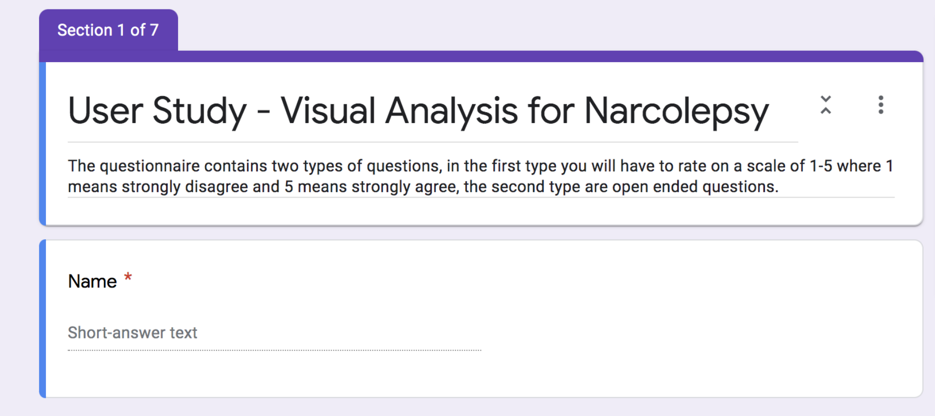
Tab 2 is called Dashboard1 - Individual and N symptoms which focuses on visualizing individual and multiple symptoms. Stacked bar plot is used to depict the distribution of symptoms present and not present and, the individual bar plots represent the distribution of symptom severity. These individual bar plots are ordered according to the order of their corresponding stacked bar plot and when the order of the stacked bar plot is changed it is reflected on these individual bar plots as well. In order to select symptoms for the parallel coordinate plots, the target user can click on either of the stacked bar plot and the parallel coordinate plots are displayed for both the subsets.

Therefore, visualization in this tab help answer the tasks, distribution of individual attributes based on symptoms present and not present, and symptom severity. It also gives the users the flexibility to choose any symptoms of their choice and visualize the selected symptoms to identify patterns, along with highlighting the importance of order.

Tab 3 is called Dashboard2 – Pairs of symptoms - this focuses on visualizing two attributes. The visualization shows the relationship among pairs of symptoms in terms of agreement and association and help identify interesting trends and groups highlighting the importance of order. The options available to interact with the heatmaps are applicable to both the population subsets. When the target user clicks on the heatmaps, corresponding balloon plots are displayed for the two population subsets.

Appendix B

User Study Questionnaire



Section 1 of 7

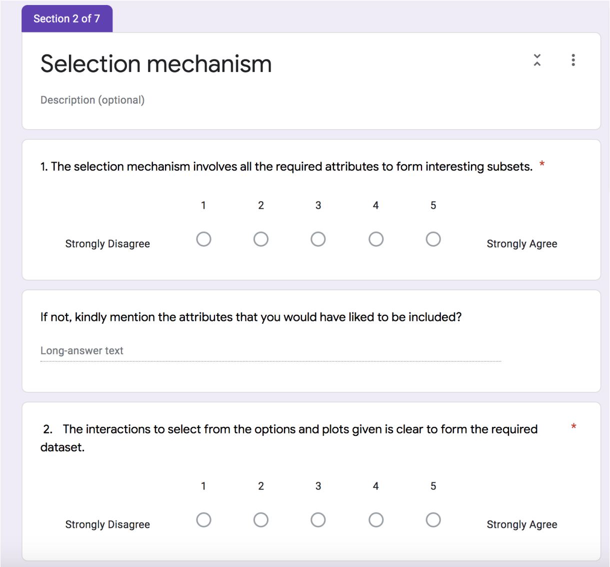
User Study - Visual Analysis for Narcolepsy

The questionnaire contains two types of questions, in the first type you will have to rate on a scale of 1-5 where 1 means strongly disagree and 5 means strongly agree, the second type are open ended questions.

Name *

Short-answer text

Figure B.1: User study questionnaire - part 1 of 11



Section 2 of 7

Selection mechanism

Description (optional)

1. The selection mechanism involves all the required attributes to form interesting subsets. *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

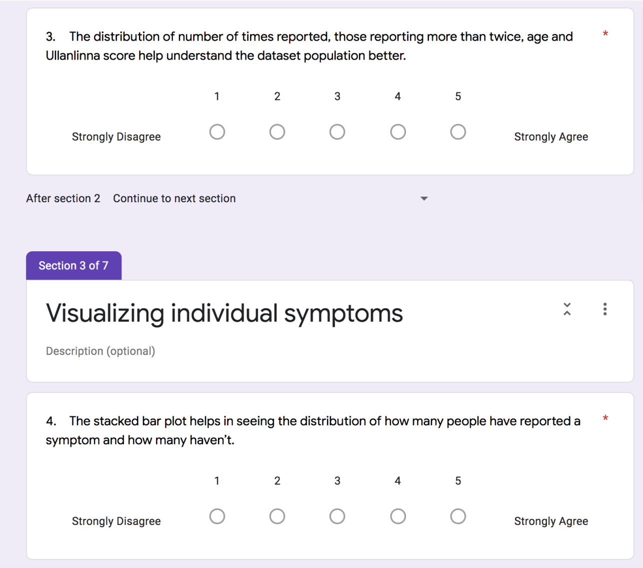
If not, kindly mention the attributes that you would have liked to be included?

Long-answer text

2. The interactions to select from the options and plots given is clear to form the required dataset. *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

Figure B.2: User study questionnaire - part 2 of 11



3. The distribution of number of times reported, those reporting more than twice, age and Ullanlinna score help understand the dataset population better. *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

After section 2 Continue to next section

Section 3 of 7

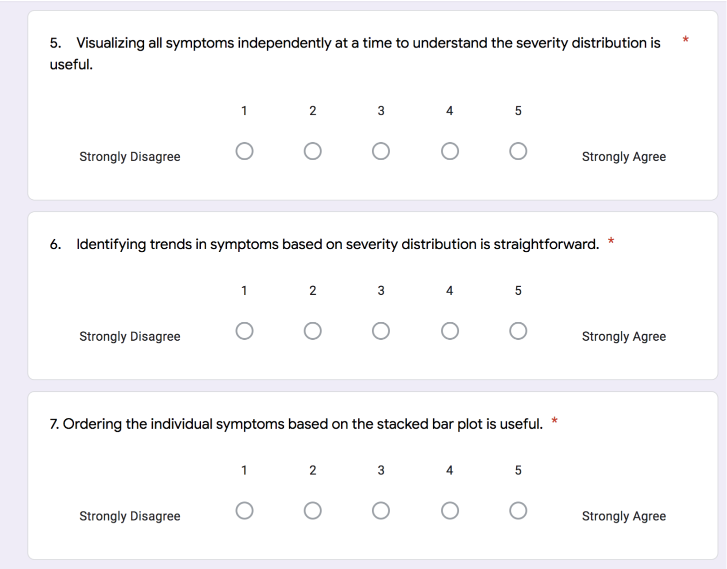
Visualizing individual symptoms

Description (optional)

4. The stacked bar plot helps in seeing the distribution of how many people have reported a symptom and how many haven't. *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

Figure B.3: User study questionnaire - part 3 of 11



5. Visualizing all symptoms independently at a time to understand the severity distribution is useful. *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

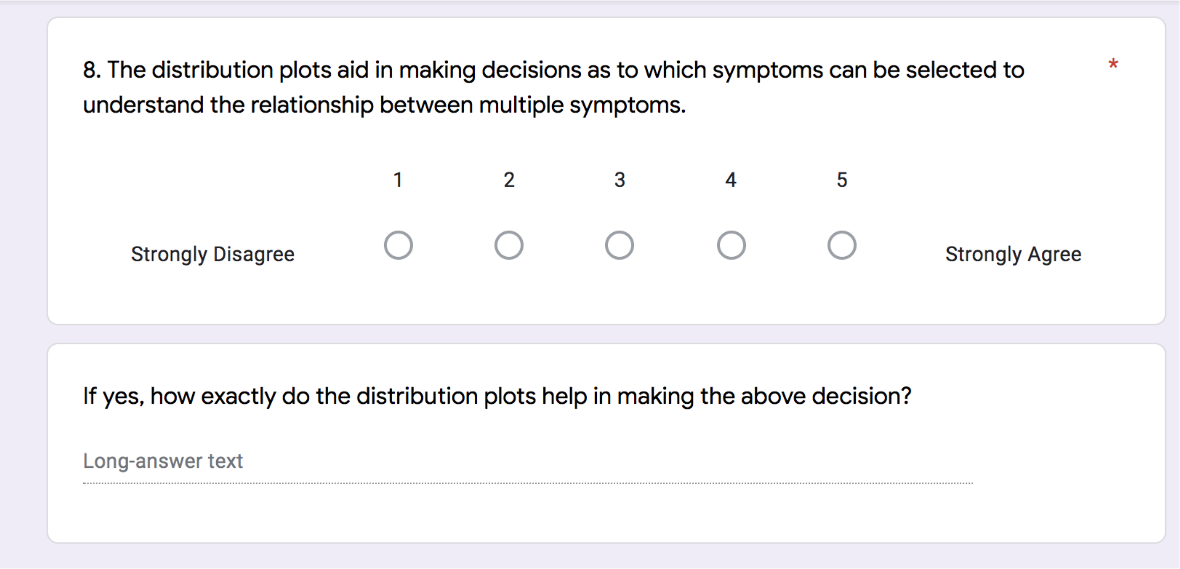
6. Identifying trends in symptoms based on severity distribution is straightforward. *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

7. Ordering the individual symptoms based on the stacked bar plot is useful. *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

Figure B.4: User study questionnaire - part 4 of 11



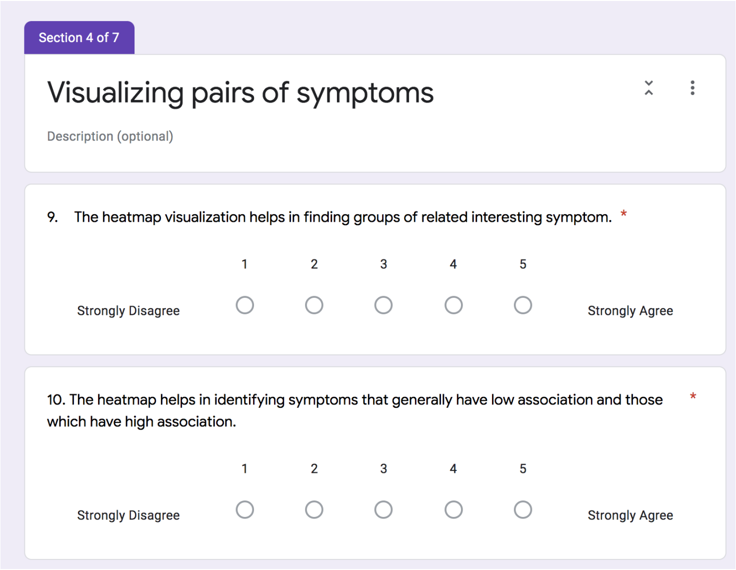
8. The distribution plots aid in making decisions as to which symptoms can be selected to understand the relationship between multiple symptoms. *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

If yes, how exactly do the distribution plots help in making the above decision?

Long-answer text

Figure B.5: User study questionnaire - part 5 of 11



Section 4 of 7

Visualizing pairs of symptoms

Description (optional)

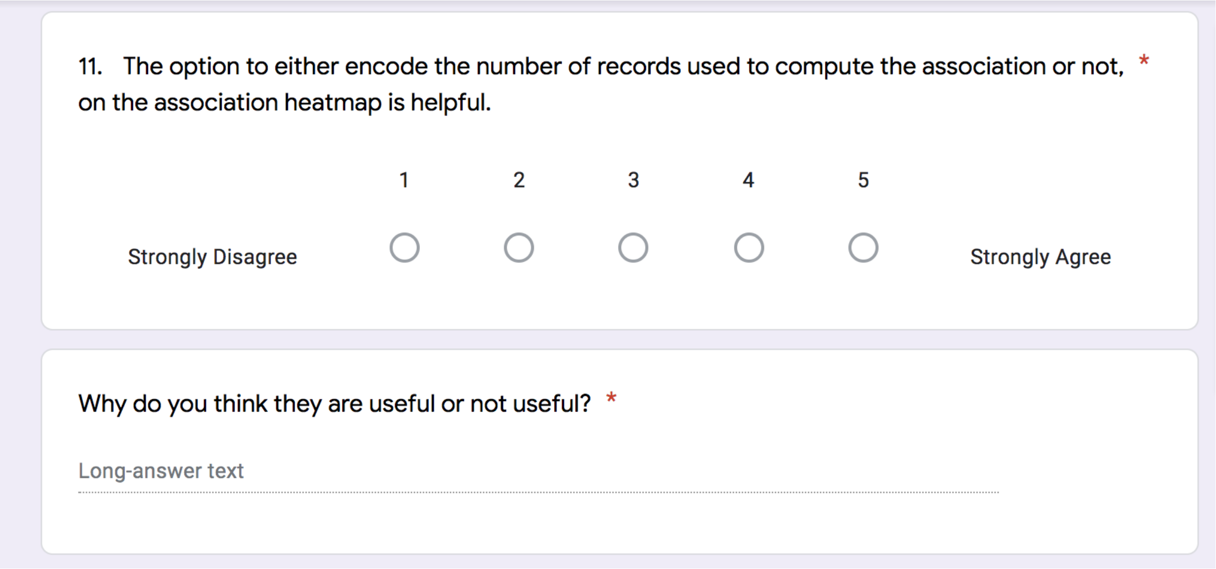
9. The heatmap visualization helps in finding groups of related interesting symptom. *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

10. The heatmap helps in identifying symptoms that generally have low association and those which have high association. *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

Figure B.6: User study questionnaire - part 6 of 11



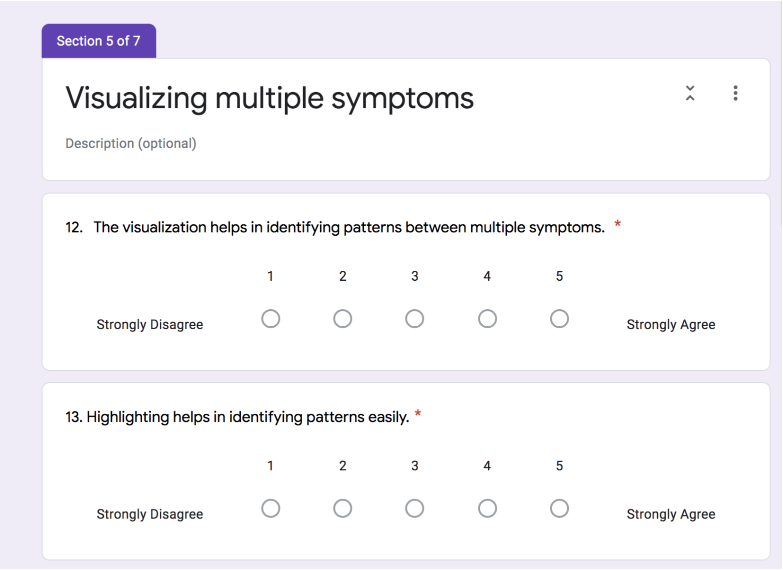
11. The option to either encode the number of records used to compute the association or not, on the association heatmap is helpful. *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

Why do you think they are useful or not useful? *

Long-answer text

Figure B.7: User study questionnaire - part 7 of 11



Section 5 of 7

Visualizing multiple symptoms

Description (optional)

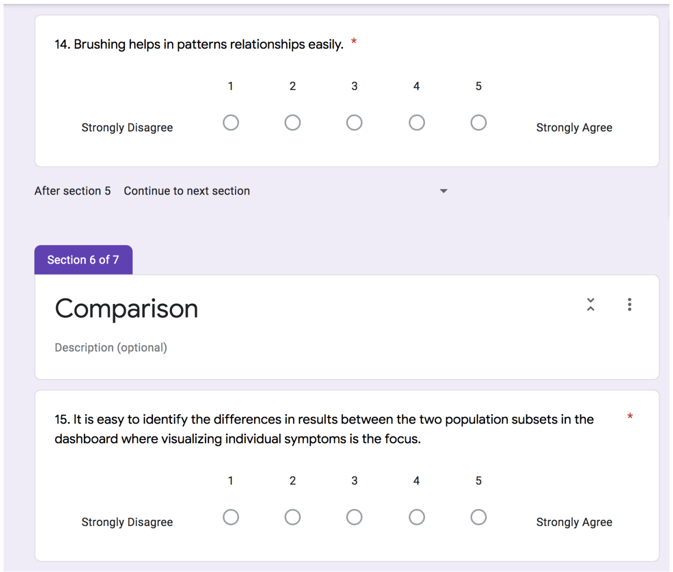
12. The visualization helps in identifying patterns between multiple symptoms. *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

13. Highlighting helps in identifying patterns easily. *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

Figure B.8: User study questionnaire - part 8 of 11



14. Brushing helps in patterns relationships easily. *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

After section 5 Continue to next section

Section 6 of 7

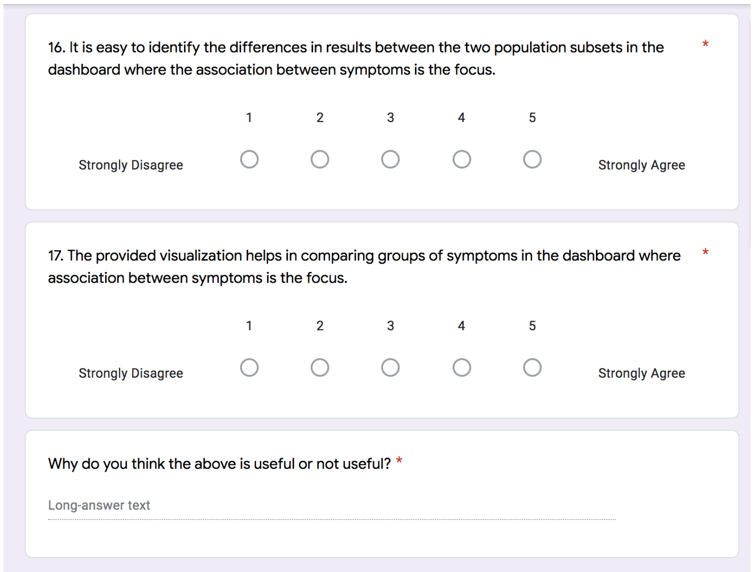
Comparison

Description (optional)

15. It is easy to identify the differences in results between the two population subsets in the dashboard where visualizing individual symptoms is the focus. *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

Figure B.9: User study questionnaire - part 9 of 11



16. It is easy to identify the differences in results between the two population subsets in the dashboard where the association between symptoms is the focus. *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

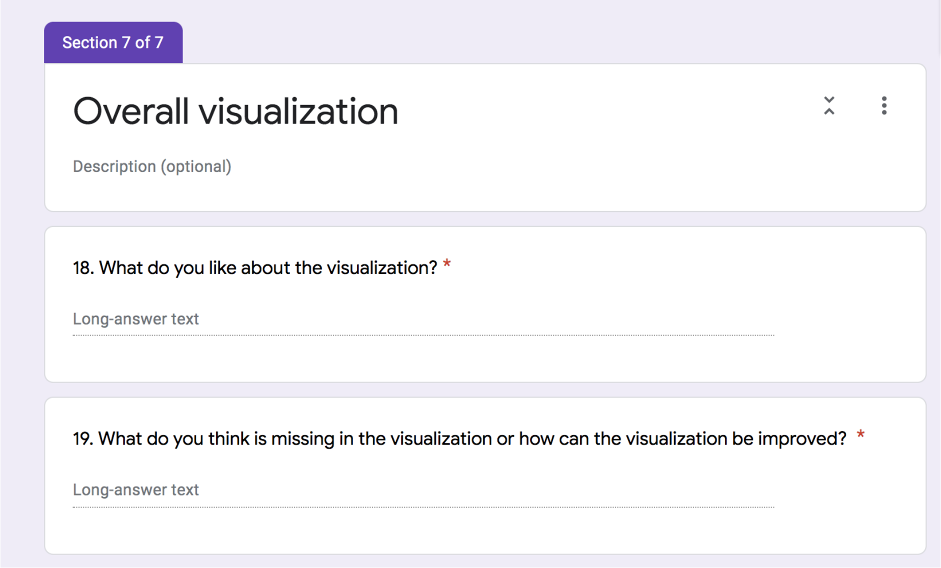
17. The provided visualization helps in comparing groups of symptoms in the dashboard where association between symptoms is the focus. *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

Why do you think the above is useful or not useful? *

Long-answer text

Figure B.10: User study questionnaire - part 10 of 11



Section 7 of 7

Overall visualization

Description (optional)

18. What do you like about the visualization? *

Long-answer text

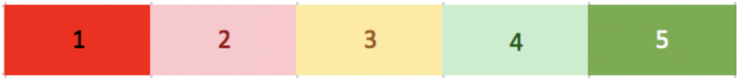
19. What do you think is missing in the visualization or how can the visualization be improved? *

Long-answer text

Figure B.11: User study questionnaire - part 11 of 11

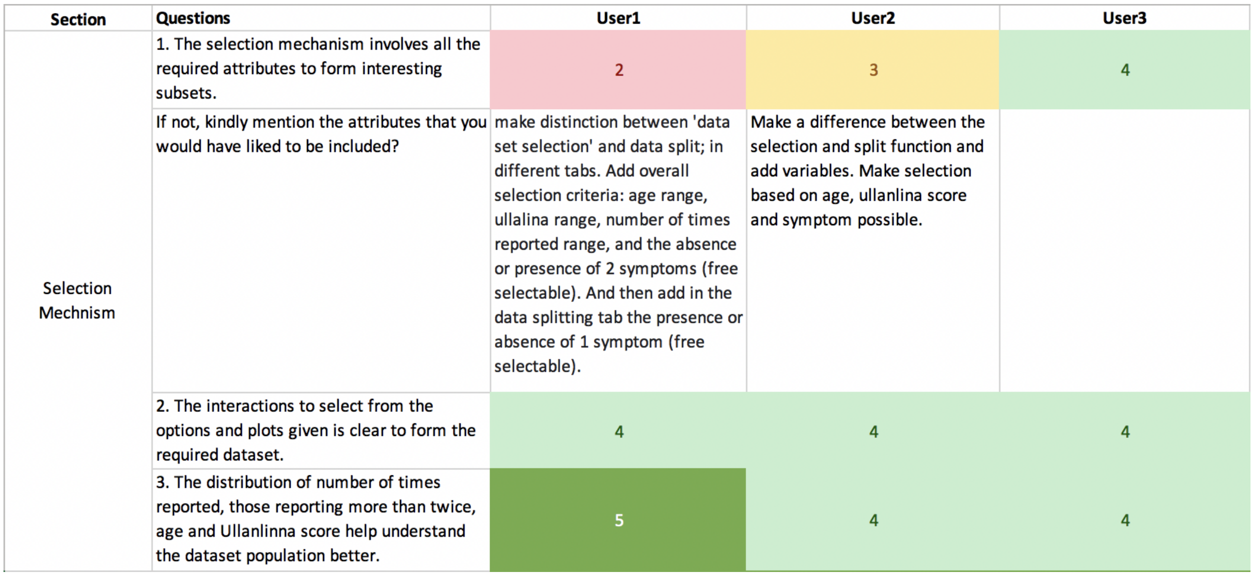
Appendix C

Complete User Study Responses



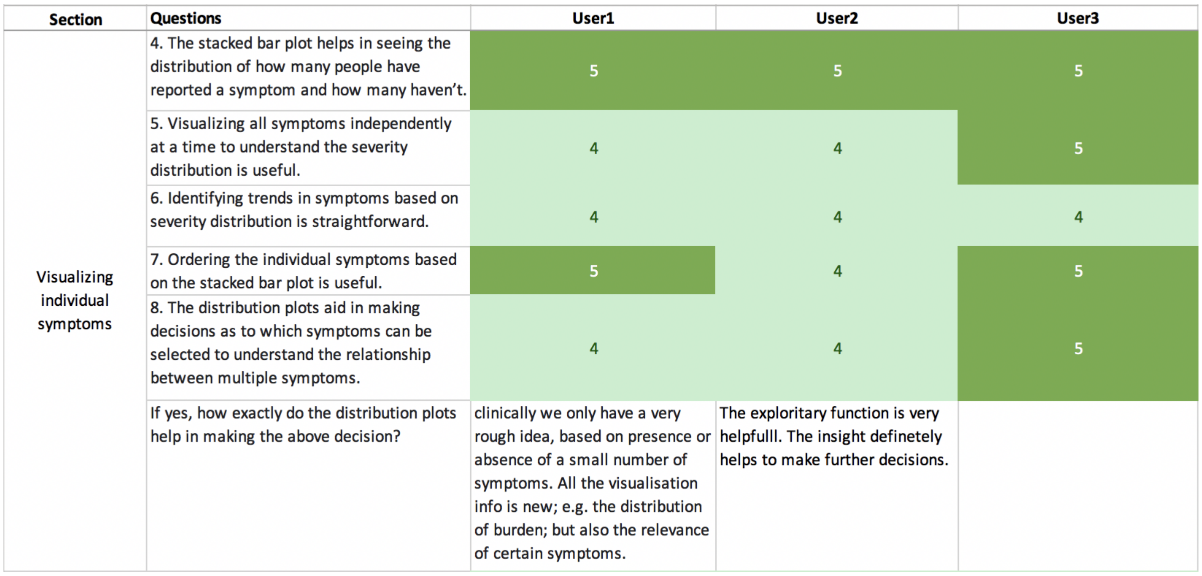
1	2	3	4	5

Figure C.1: Key for user study responses



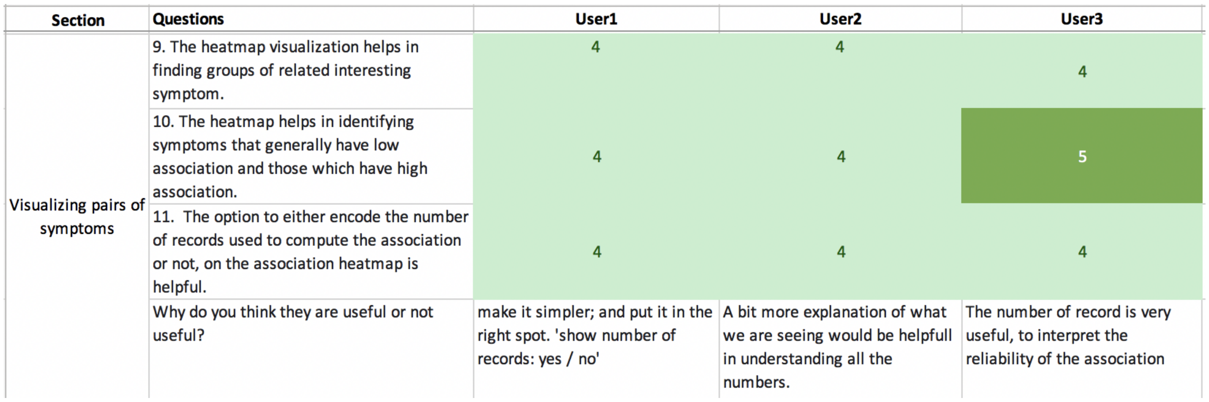
Section	Questions	User1	User2	User3
Selection Mechnism	1. The selection mechanism involves all the required attributes to form interesting subsets.	2	3	4
	If not, kindly mention the attributes that you would have liked to be included?	make distinction between 'data set selection' and data split; in different tabs. Add overall selection criteria: age range, ullalina range, number of times reported range, and the absence or presence of 2 symptoms (free selectable). And then add in the data splitting tab the presence or absence of 1 symptom (free selectable).	Make a difference between the selection and split function and add variables. Make selection based on age, ullanlina score and symptom possible.	
	2. The interactions to select from the options and plots given is clear to form the required dataset.	4	4	4
	3. The distribution of number of times reported, those reporting more than twice, age and Ullanlinna score help understand the dataset population better.	5	4	4

Figure C.2: User Study Responses - Selection Mechanism



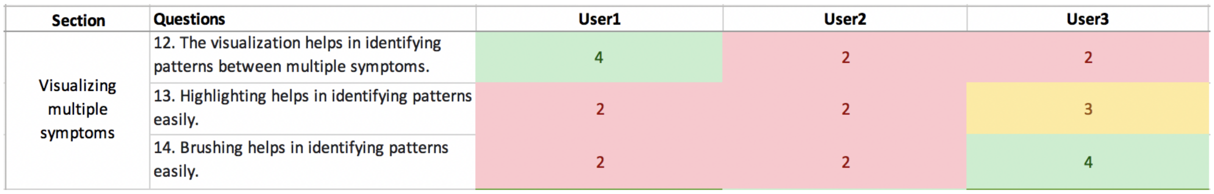
Section	Questions	User1	User2	User3
Visualizing individual symptoms	4. The stacked bar plot helps in seeing the distribution of how many people have reported a symptom and how many haven't.	5	5	5
	5. Visualizing all symptoms independently at a time to understand the severity distribution is useful.	4	4	5
	6. Identifying trends in symptoms based on severity distribution is straightforward.	4	4	4
	7. Ordering the individual symptoms based on the stacked bar plot is useful.	5	4	5
	8. The distribution plots aid in making decisions as to which symptoms can be selected to understand the relationship between multiple symptoms.	4	4	5
	If yes, how exactly do the distribution plots help in making the above decision?	clinically we only have a very rough idea, based on presence or absence of a small number of symptoms. All the visualisation info is new; e.g. the distribution of burden; but also the relevance of certain symptoms.	The exploritary function is very helpfulll. The insight definetely helps to make further decisions.	

Figure C.3: User Study Responses - Visualizing individual symptoms



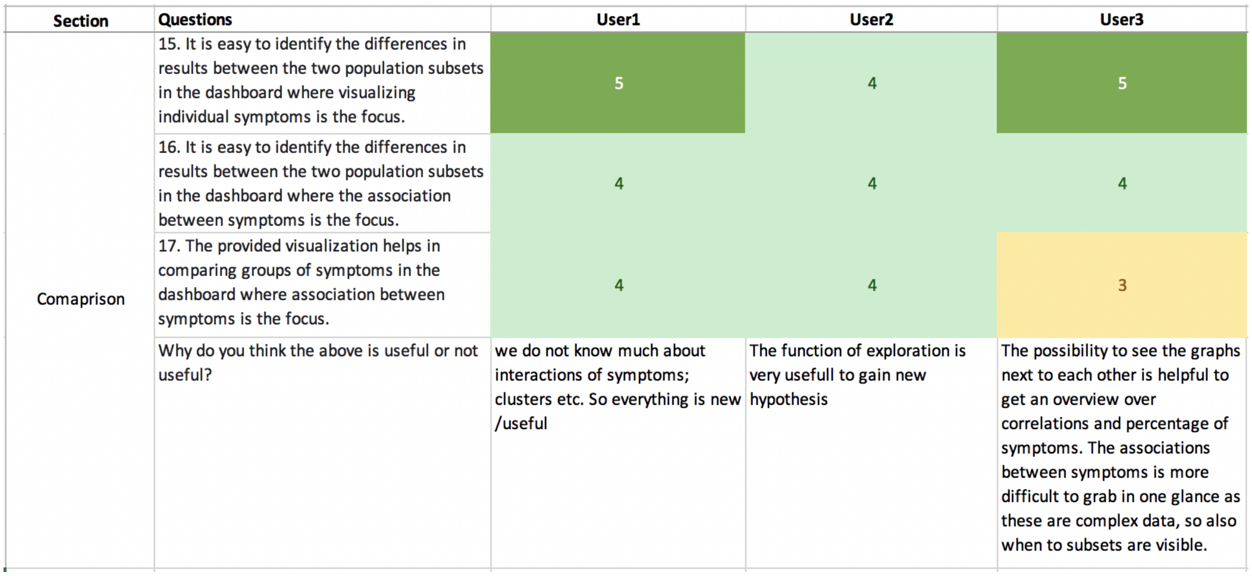
Section	Questions	User1	User2	User3
Visualizing pairs of symptoms	9. The heatmap visualization helps in finding groups of related interesting symptom.	4	4	4
	10. The heatmap helps in identifying symptoms that generally have low association and those which have high association.	4	4	5
	11. The option to either encode the number of records used to compute the association or not, on the association heatmap is helpful.	4	4	4
	Why do you think they are useful or not useful?	make it simpler; and put it in the right spot. 'show number of records: yes / no'	A bit more explanation of what we are seeing would be helpfull in understanding all the numbers.	The number of record is very useful, to interpret the reliability of the association

Figure C.4: User Study Responses - Visualizing pairs of symptoms



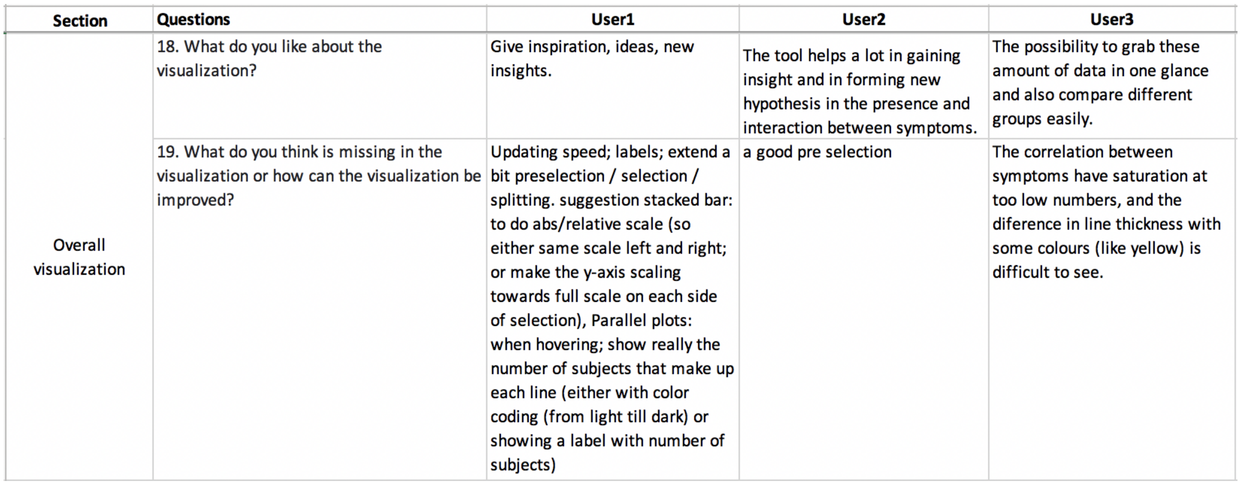
Section	Questions	User1	User2	User3
Visualizing multiple symptoms	12. The visualization helps in identifying patterns between multiple symptoms.	4	2	2
	13. Highlighting helps in identifying patterns easily.	2	2	3
	14. Brushing helps in identifying patterns easily.	2	2	4

Figure C.5: User Study Responses - Visualizing multiple symptoms



Section	Questions	User1	User2	User3
Comaprison	15. It is easy to identify the differences in results between the two population subsets in the dashboard where visualizing individual symptoms is the focus.	5	4	5
	16. It is easy to identify the differences in results between the two population subsets in the dashboard where the association between symptoms is the focus.	4	4	4
	17. The provided visualization helps in comparing groups of symptoms in the dashboard where association between symptoms is the focus.	4	4	3
	Why do you think the above is useful or not useful?	we do not know much about interactions of symptoms; clusters etc. So everything is new /useful	The function of exploration is very usefull to gain new hypothesis	The possibility to see the graphs next to each other is helpful to get an overview over correlations and percentage of symptoms. The associations between symptoms is more difficult to grab in one glance as these are complex data, so also when to subsets are visible.

Figure C.6: User Study Responses - Comparison



Section	Questions	User1	User2	User3
Overall visualization	18. What do you like about the visualization?	Give inspiration, ideas, new insights.	The tool helps a lot in gaining insight and in forming new hypothesis in the presence and interaction between symptoms.	The possibility to grab these amount of data in one glance and also compare different groups easily.
	19. What do you think is missing in the visualization or how can the visualization be improved?	Updating speed; labels; extend a bit preselection / selection / splitting. suggestion stacked bar: to do abs/relative scale (so either same scale left and right; or make the y-axis scaling towards full scale on each side of selection), Parallel plots: when hovering; show really the number of subjects that make up each line (either with color coding (from light till dark) or showing a label with number of subjects)	a good pre selection	The correlation between symptoms have saturation at too low numbers, and the diference in line thickness with some colours (like yellow) is difficult to see.

Figure C.7: User Study Responses - Overall visualization

Bibliography

[1] Birgitte R. Kornum, Stine Knudsen, Hanna M. Ollila, Fabio Pizza, Poul J. Jennum, Yves Dauvilliers, and Sebastiaan Overeem. *Narcolepsy*, volume 3. 2017. ISBN 2056-676X. doi: 10.1038/nrdp.2016.100. URL `https://doi.org/10.1038/nrdp.2016.100`.

[2] Chand M. Ruoff, Nancy L. Reaven, Susan E. Funk, Karen J. McGaughey, Maurice M. Ohayon, Christian Guilleminault, and Jed Black. *High Rates of Psychiatric Comorbidity in Narcolepsy: Findings From the Burden of Narcolepsy Disease (BOND) Study of 9,312 Patients in the United States.*, volume 78(2):171-176. 2017. doi: 10.4088/JCP.15m10262. URL `https://doi.org/10.4088/jcp.15m10262`.

[3] Y. Dauvilliers, I. Arnulf, and E. Mignot. *Narcolepsy with cataplexy*, volume 369, 9560: 499-511. 2007. doi: 10.1016/S0140-6736(07)60237-2. URL `https://doi.org/10.1016/S0140-6736(07)60237-2`.

[4] Michael J. Thorpy and Ana C. Krieger. *Delayed diagnosis of narcolepsy: characterization and impact*, volume 15,5: 502-507. 2014. doi: 10.1016/j.sleep.2014.01.015. URL `https://doi.org/10.1016/j.sleep.2014.01.015`.

[5] Maurice M. Ohayon. *Narcolepsy is complicated by high medical and psychiatric comorbidities: a comparison with the general population*, volume 14,6: 488-492. 2013. doi: 10.1016/j.sleep.2013.03.002. URL `https://doi.org/10.1016/j.sleep.2013.03.002`.

[6] Michael Thorpy and Anne Marie Morse. *Reducing the Clinical and Socioeconomic Burden of Narcolepsy by Earlier Diagnosis and Effective Treatment*, volume 12,1: 61-71. 2017. doi: 10.1016/j.jsmc.2016.10.001. URL `https://doi.org/10.1016/j.jsmc.2016.10.001`.

[7] C. Hublin, J. Kaprio, M. Partinen, M. Koskenvuo, and K. Heikkilä. *The Ullanlinna Narcolepsy Scale: validation of a measure of symptoms in the narcoleptic syndrome*, volume 3,1: 52-59. 1994. doi: 10.1111/j.1365-2869.1994.tb00104.x. URL `https://doi.org/10.1111/j.1365-2869.1994.tb00104.x`.

[8] Sarah Faisal, Ann Blandford, and Henry WW Potts. *Making sense of personal health information: Challenges for information visualization*, volume 19,3. 2013. doi: 10.1177/1460458212465213. URL `https://doi.org/10.1177/1460458212465213`. PMID: 23981395.

[9] D. L. Gresh, D. A. Rabenhorst, A. Shabo, and S. Slavin. *PRIMA: A case study of using information visualization techniques for patient record analysis.* Oct 2002. ISBN 0-7803-7498-3. doi: 10.1109/VISUAL.2002.1183817. URL `https://doi.org/10.1109/VISUAL.2002.1183817`.

[10] Michael Spenke. *Visualization and Interactive Analysis of Blood Parameters with InfoZoom*, volume 22,2. Elsevier Science Publishers Ltd., GBR, May 2001. doi: 10.1016/S0933-3657(00)00105-6. URL `https://doi.org/10.1016/S0933-3657(00)00105-6`.

[11] Taowei David Wang, Catherine Plaisant, Alexander J. Quinn, Roman Stanchak, Shawn Murphy, and Ben Shneiderman. *Aligning Temporal Data by Sentinel Events: Discovering Patterns in Electronic Health Records.* CHI '08. Association for Computing Machinery,

New York, NY, USA, 2008. ISBN 9781605580111. doi: 10.1145/1357054.1357129. URL https://doi.org/10.1145/1357054.1357129.

[12] Taowei David Wang, Krist Wongsuphasawat, Catherine Plaisant, and Ben Shneiderman. *Extracting Insights from Electronic Health Records: Case Studies, a Visual Analytics Process Model, and Design Recommendations*, volume 35. 2011. ISBN 1573-689X. doi: 10.1007/s10916-011-9718-x. URL https://doi.org/10.1007/s10916-011-9718-x.

[13] Krist Wongsuphasawat and David Gotz. *Outflow: Visualizing Patient Flow by Symptoms and Outcome*. 01 2011.

[14] Susana B Martins, Yuval Shahar, Maya Galperin, Herbert Kaizer, Dina Goren-Bar, Deborah McNaughton, Lawrence V Basso, and Mary K Goldsteinc. *Evaluation of KNAVE-II: a tool for intelligent query and exploration of patient data*, volume 107,1,648-652. 2004. doi: 10.3233/978-1-60750-949-3-648.

[15] Dina Goren-Bar, Yuval Shahar, Maya Galperin-Aizenberg, David Boaz, and Gil Tahan. *KNAVE II: The Definition and Implementation of an Intelligent Tool for Visualization and Exploration of Time-Oriented Clinical Data*. AVI '04. Association for Computing Machinery, New York, NY, USA, 2004. ISBN 1581138679. doi: 10.1145/989863.989889. URL https://doi.org/10.1145/989863.989889.

[16] Yuval Shahar, Dina Goren Bar, Maya Galperin-Aizenberg, David Boaz, and Gil Tahan. *KNAVE-II: A distributed architecture for interactive visualization and intelligent exploration of time-oriented clinical data*. 01 2003.

[17] Yuval Shahar, Dina Goren Bar, David Boaz, and Gil Tahan. *Distributed, intelligent, interactive visualization and exploration of time-oriented clinical data and their abstractions*, volume 38. 11 2006. doi: 10.1016/j.artmed.2005.03.001. URL https://doi.org/10.1016/j.artmed.2005.03.001.

[18] S. Zillner, T. Hauer, D. Rogulin, A. Tsymbal, M. Huber, and T. Solomonides. *Semantic Visualization of Patient Information*. June 2008. doi: 10.1109/CBMS.2008.11. URL https://doi.org/10.1109/CBMS.2008.11.

[19] Krist Wongsuphasawat, John Alexis Guerra Gómez, Catherine Plaisant, Taowei David Wang, Meirav Taieb-Maimon, and Ben Shneiderman. *LifeFlow: Visualizing an Overview of Event Sequences*. CHI '11. Association for Computing Machinery, New York, NY, USA, 2011. ISBN 9781450302289. doi: 10.1145/1978942.1979196. URL https://doi.org/10.1145/1978942.1979196.

[20] Wei Chen, Fangzhou Guo, and Fei-Yue Wang. *A Survey of Traffic Data Visualization*, volume 16. 2015.

[21] H. Guo, Z. Wang, B. Yu, H. Zhao, and X. Yuan. *TripVista: Triple Perspective Visual Trajectory Analytics and its application on microscopic traffic data at a road intersection*. March 2011. doi: 10.1109/PACIFICVIS.2011.5742386.

[22] Tamara Munzner. *Visualization Analysis & Design*. A K Peters/CRC Press, 2014. ISBN 9781498759717.

[23] Harry Khamis. *Measures of Association How to Choose?*, volume 24,3. JDMS, 2008.

[24] Robert Gove, Nick Gramsky, Rose Kirby, Emre Sefer, Awalin Sopan, Cody Dunne, Ben Shneiderman, and Meirav Taieb-Maimon. *NetVisia: Heat Map & Matrix Visualization of Dynamic Social Network Statistics & Content*.

[25] Danielle Albers Szafir. *The Good, the Bad, and the Biased: Five Ways Visualizations Can Mislead (and How to Fix Them)*. 2018.

[26] Simone Kriglstein, Margit Pohl, and Michael Smuc. *Pep Up Your Time Machine: Recommendations for the Design of Information Visualizations of Time-Dependent Data.* Springer New York, New York, NY, 2014. ISBN 978-1-4614-7485-2. doi: 10.1007/978-1-4614-7485-2_8. URL `https://doi.org/10.1007/978-1-4614-7485-2_8`.