Otto-von-Guericke University Magdeburg

Faculty of Computer Science
Department of Simulation and Graphics

Dissertation

# Visual Analytics of Epidemiological and Multi-Omics Data

**Author:**
Shiva Alemzadeh
Born on June 22, 1987 in Shiraz

**Reviewers:**
Prof. Dr. Bernhard Preim
Prof. Dr. Silvia Miksch
Prof. Dr. Jörn Kohlhammer
May 5, 2021

**Alemzadeh, Shiva:**
*Visual Analytics of Epidemiological and Multi-Omics Data*
Dissertation, Otto-von-Guericke University
Magdeburg, 2020.

# Zusammenfassung

Kohortenstudien zielen darauf ab, Risikofaktoren zu identifizieren, die den Gesundheitszustand einer Bevölkerung in Bezug auf bestimmte Krankheiten beeinflussen können. In der Epidemiologie werden die Teilnehmer von Gesundheitsstudien zur Ermittlung der Risikofaktoren einer Krankheit hinsichtlich verschiedener gesundheitsbezogener Aspekte beobachtet, die bei der Entwicklung der Zielkrankheit eine Rolle spielen können. Diese Aspekte umfassten viele Faktoren wie Lebensstil (z. B. Rauchen oder Alkoholkonsum) oder Medikamente (z. B. Drogenkonsum), aktuelle Gesundheit (z. B. Diabetes oder erhöhter Blutdruck) und soziodemografische Faktoren (z. B. Geschlecht, Familienstand).
Um die Daten zu sammeln, werden Personen eingeladen, an der Studie teilzunehmen. Die Informationen von den Individuen werden durch Interviews erhalten, z.B. Fragen zu ihren Gewohnheiten und dem aktuellen Gesundheitszustand. Darüber hinaus werden medizinische Bilder akquiriert, um Anomalien zu entdecken, z.B. erhöhte Brustdichte. In Kohortenstudien werden dieselben Personen erneut zur Studie eingeladen, um die Untersuchungen zu wiederholen und die zeitlichen Veränderungen zu beobachten. Der Zweck dieser Nachuntersuchungen besteht darin, den möglichen Zusammenhang zwischen den Krankheiten und den Risikofaktoren zu identifizieren. Fehlende Daten sind jedoch ein unvermeidlicher Bestandteil solcher Studien, bei denen einige Personen aus der Studie ausscheiden oder die Aufzeichnungen unvollständig sind. Diese Arbeit bietet halbüberwachte visuelle Analyse-Frameworks, um diskriminierende Subpopulationen in Kohortenstudiendaten zu untersuchen und zu entdecken. Um dies zu erreichen, bieten interaktiv koordinierte Mehrfachansichtsysteme die Möglichkeit, um die verschiedenen Assoziationen zwischen Daten und Funktionen zu untersuchen. Mit S-ADVIsED kann der Analyst die Ergebnisse des Subspace-Clusters untersuchen und visuell validieren. Mit DiscoVA kann der Analyst Subpopulationen anhand verschiedener Datentypen (z. B.Multi-Omics und klinische Daten) identifizieren.
Um fehlende Daten in Längsschnittstudiendaten zu behandeln, bietet das VIVID Framework außerdem Methoden zur Imputation (z. B. Mehrfachimputation) und zur Überprüfung der Plausibilität der Ergebnisse.

## Abstract

Cohort studies aim to identify risk factors that may influence the health conditions of a population regarding specific diseases. In epidemiology science, to discover the risk factors of a disease, the cohort individuals are observed regarding different health-related aspects that may have a role in developing the target disease. These aspects involved many factors like lifestyle (e.g. smoking or alcohol consumption), or medicament (e.g. taking drugs), current health situations (e.g. diabetes or having elevated blood pressure), and socio-demographic factors (e.g. gender, marital status).

To collect the data, people are invited to join the study via advertisements. The information from the individuals is acquired by having interviews, e.g. asking questions about their habits examinations and doing to find out the current health condition of individuals. Moreover, medical images are prepared to discover probable abnormalities, e.g. breast density. In cohort study data, the same individuals are re-invited to the study to repeat the examinations and observe the changes over time in different time points. The purpose of these follow-ups is to identify the possible link between the diseases and risk factors. However, missing data are an inevitable part of such studies where some individuals drop from the study of the records is incomplete. This thesis provides semi-supervised visual analytics frameworks to explore and discover discriminative subpopulations in cohort study data. To reach this, interactive coordinated multiple views systems provide a platform to investigate the different associations between the data and features. S-ADVIsED enables the analyst to explore the results of subspace clustering and also validate the results visually. DiscoVA enables the analyst to identify subpopulations using different data types (e.g. multi-omics and clinical data).

Additionally, to handle missing data in longitudinal study data, VIVID framework provides the methods to explore and impute (e.g. multiple imputation) the missing values. Moreover, it allows the expert to check the plausibility of the results.

# Contents

**Zusammenfassung**

**Abstract**

**1 Introduction and Motivation**

**2 Background & Methods**
2.1 Epidemiological Study Terms and Aims . . . . . . . . . . . . . . . . . 6
2.1.1 Cohort Study Data Sources . . . . . . . . . . . . . . . . . . . . . 9
2.1.2 Cross-Sectional Study vs. Longitudinal Study . . . . . . . . . 10
2.1.3 The Role of Data Mining in Epidemiological Studies . . . . . 11
2.2 Study Examples . . . . . . . . . . . . . . . . . . . . . . . . . . . . 30

**3 Data Quality**
3.1 Missing Values as a Data quality Issue . . . . . . . . . . . . . . . . . 34
3.2 Related Works in VA . . . . . . . . . . . . . . . . . . . . . . . . . . 37
3.2.1 Handling Missing Epidemiological Data . . . . . . . . . . . . 38
3.2.2 Task Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . 40
3.2.3 VIVID . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 41
3.2.4 VIVID Workflow . . . . . . . . . . . . . . . . . . . . . . . . . 48
3.3 Evaluation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 51
3.3.1 Quality of Multiple Imputation . . . . . . . . . . . . . . . . . 51
3.3.2 Trade-off between Accuracy and Time . . . . . . . . . . . . . 55
3.3.3 Summary and Conclusion . . . . . . . . . . . . . . . . . . . . 56

**4 Epidemiological Sub-cohort Discovery**
4.1 Clustering . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 64
4.2 Related Works in VA . . . . . . . . . . . . . . . . . . . . . . . . . . 67
4.3 Data for Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . 70
4.4 Subpopulation Discovery in Cross-Sectional Data . . . . . . . . . . . 72
4.4.1 Task Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . 72
4.4.2 S-ADVIsED System . . . . . . . . . . . . . . . . . . . . . . . . 74
4.5 Subpopulation Discovery in Longitudinal Data . . . . . . . . . . . . 82
4.5.1 Task analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . 83
4.5.2 Mock-ups . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 84

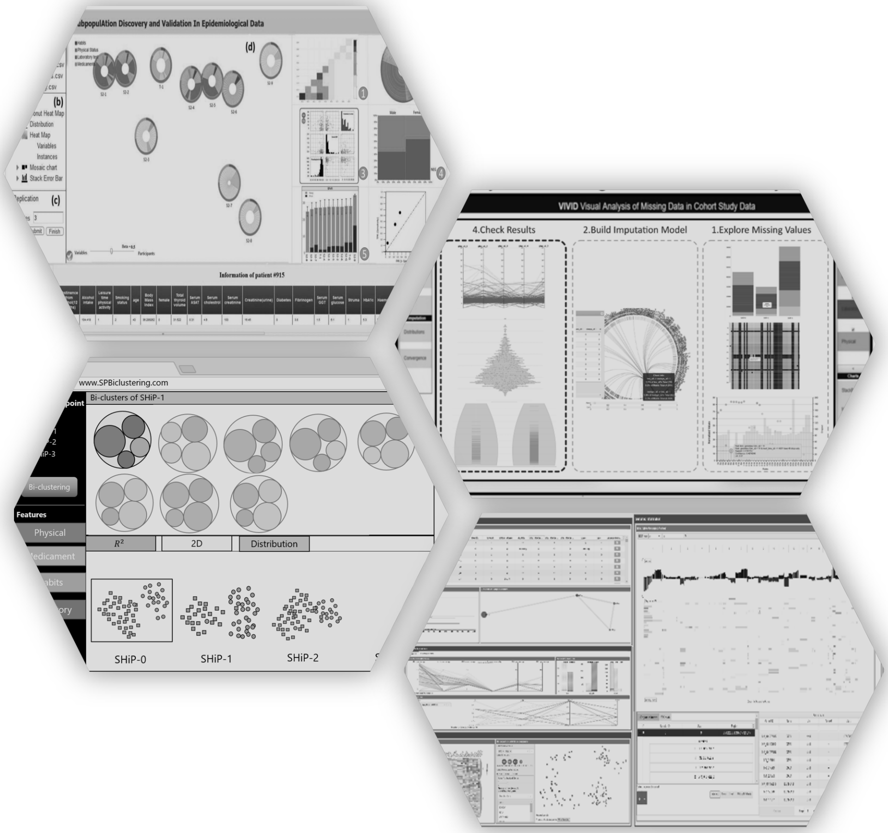
4.6 Summary and Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . 86

**5 Sub-cohort Discovery of Multi-Omics Data**
5.1 Related Works of VA . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 92
5.2 Background and Data . . . . . . . . . . . . . . . . . . . . . . . . . . . . 94
5.3 Limitations . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 97
5.4 DiscoVA system . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 98
  5.4.1 System Design and Task Analysis . . . . . . . . . . . . . . . . . . 98
  5.4.2 Components . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 100
5.5 Evaluation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 110
  5.5.1 Case Study . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 111
  5.5.2 Survey . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 113
5.6 Summary and Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . 114

**6 Conclusions**
6.1 Visual Analytics of Missing Values . . . . . . . . . . . . . . . . . . . . 118
6.2 Cross-Sectional & Longitudinal Epidemiological Data . . . . . . . . 119
6.3 Visual Queries on Multi-Omics Data . . . . . . . . . . . . . . . . . . . 120

# 1

# Introduction and Motivation



The main aim of population health studies as an interdisciplinary field is to improve the health situation of a group of people who have some shared characteristics, e.g. they reside in the same area, they have the same lifestyle or they are of the same ethnicity [1]. The experts of population health investigate the determinants which influence the level of health within and across populations. The factors that may influence the determinants consist of the biological history, lifestyle, and socio-demographic status of a population.

In this thesis, we try to find an answer to the question of what level of guidance is necessary to support the biomedical experts for analysis of the data? and how the quality of the data may influence the analyses? Additionally, is the analysis of large-scale with divergence types possible? There are certain obstacles to make a complex analysis of medical data possible for medical experts. Firstly, for analysis of the data, the biomedical experts usually do not have enough programming skills and prefer not to be thrown into a pool of parameters because it costs a lot of time and effort to find out the influence of changing each parameter on the result. Secondly, the medical data are high dimensional and heterogeneous, thus it makes it hard to aggregate the data and find answers for the medical questions.

Generally, this thesis tackles the following challenges:

- Improving data quality using visualization techniques along with statistical and data mining techniques on longitudinal epidemiological studies
- Allowing biomedical experts to derive and validate complex hypotheses
- Introducing semi-exploitative approaches to analysis complex heterogeneous medical data

To find patterns among the epidemiological data we need to analyze them by statistical or machine learning techniques. Before performing any technique we should make sure that the results are trustable. Thus, to achieve correct results we need to consider the correctness of the **data** and the applied **techniques**.

Thus, a part of this thesis focuses on data quality, i.e. missing data issues as an inevitable part of healthcare studies. The remaining part of this thesis provides a combination of data mining techniques as well as visualizations to find a pattern on the data, for example, identifying high-risk sub-cohorts vs. low-risk sub-cohorts.

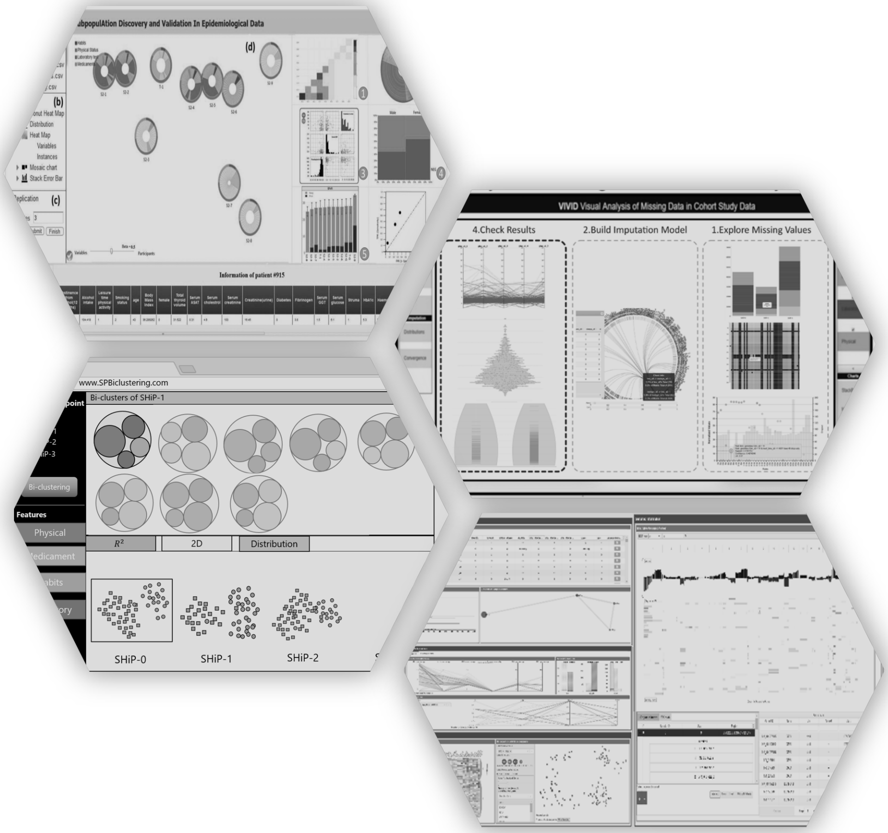
This thesis is organized as follows:

- Chapter 2 provides information on the basis of epidemiological studies including the history, data, and aims. Moreover, a brief description of the widely used machine learning techniques for the analysis of epidemiological data is given.
- Chapter 3 covers a summary of the data quality issues. The main focus of this chapter are missing data issues in longitudinal epidemiological data. A proposed framework for handling missing data using visualization techniques is explained.
- Chapter 4 covers the sub-cohort discovery in epidemiological data. The clustering techniques are used as the core of this chapter. A proposed framework for sub-cohort validation and discovery in cross-sectional data is described. Additionally, the mock-ups for the extension work for sub-cohort discovery in longitudinal data are provided.
- Chapter 5 focuses on the sub-cohort discovery of multi-omics data. The provided framework for making complex visual queries on cancer patients' heterogeneous data types and sources is described.
- Finally, Chapter 6 concludes the thesis by summarizing the contributions of the thesis.

Chapters 3,4 and 5 are based on conference and journal publications.

# 2

# Background & Methods



In this chapter, a general description of epidemiology aims, terms, data, and techniques to analyze the data is provided. The main focus of this thesis is employing data mining techniques to find patterns among epidemiological data. Thus, the data mining techniques consisting of clustering, classification, feature selection, and dimension reduction techniques are described more specifically.

## 2.1 Epidemiological Study Terms and Aims

Epidemiology studies refer to the occurrence and reasons of disease in specific populations. The aims of such studies are to prevent diseases by creating strategies using investigations in populations who already experienced a disease.

Epidemiological studies measure the characteristics of populations. The parameter of interest may be a disease rate, the prevalence of exposure, or more often some measure of the association between an exposure and disease. Because studies are carried out on people and have all the attendant practical and ethical constraints, they are almost invariably subject to bias.

The epidemiology science is about 2500 years old. Following gives an introduction on epidemiology emergence and evolution [2]:

- **Hippocrates:** Hippocrates was a Greek physician and the first epidemiologist in history. He is known as "the father of medicine". He was the first person who made a distinction between the terms "epidemic" and "endemic". When a special disease exists permanently or for a long time in a region, it is called an "endemic" disease, for example, the HIV infection is an "endemic" in parts of Africa. The high prevalence of a disease in a high number of people at the same time and in the same community is called an "epidemic". The epidemic diseases may spread through communities. When the disease spreads all over the world, its disease is called a "pandemic". The COVID-19 is an example of a pandemic disease in 2020 [3].
- **John Graunt:** John Graunt was a haberdasher and councilman in London. In 1662, he published a landmark on the analysis of mortality data. His contribution to epidemiology leads to understand-

ing the facts on human life and diseases. One of his contributions was the discovery that the ratio of births and deaths of men is more than that of women, namely a ratio of 14 to 13. He also noticed that although women have lower mortality, they get ill more often than men.

- **William Farr:** William Farr was a mortality statistician from Britain and is known as the father of modern vital statistics. He introduced many basic statistical models that are widely used today in the classification of diseases.
- **John Snow:** In 1854, and in the time of exploitation of the cholera epidemic in the Golden Square of London, John Snow made some investigations to discover the cause of the disease and find solutions to prevent it. He believed that the source of cholera infection was the water. Thus, he illustrated a map of the Golden Square area to make hypotheses about the source of cholera including the water pumps, see Figure 2.1. In his illustration, called spot map, he marked the geographical distribution of cholera cases and the water pumps to find out the relationship between residents with cholera and water pumps. In the end, he concluded that the Broad Street pump was the main source of infection [4].

In the late 1800s, using epidemiological studies for investigation of infectious disease prevalence became more common. Later, between 1930 and 1940, epidemiological research was extended to noninfectious diseases. Afterward, epidemiology science was applied to almost all health-related issues.

**Epidemiological Data Types.** In this section, a description of sources and types of epidemiological data is provided. Generally, based on the data collection, the epidemiological studies are classified into two categories the observational or experimental. In the observational type, a disease like a flue is analyzed in the wild, while in the experimental type the expert controls the study features, e.g. drug trials [5]. The main focus of this thesis is based on observational studies. The epidemiological studies can be categorized into non-inclusive three stages/types. Figure 2.2 shows a taxonomy of the cohort studies. In the following, the observational studies are described [6] and [7].

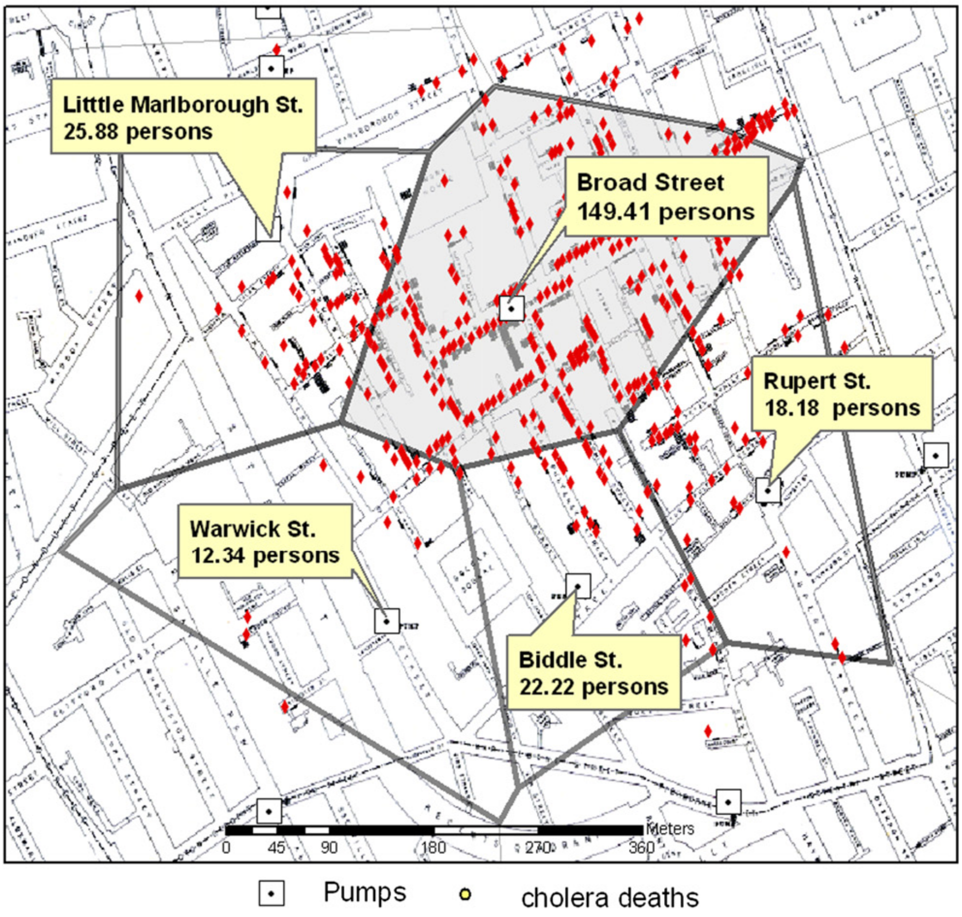


Figure 2.1: John Snow on cholera. London: Humphrey Milford: Oxford University Press; 1936.

**Cohort studies:** The cohort studies are usually helpful to understand the pattern of spreading of diseases in a population. The results can be extended to wider populations. These studies are mostly used for preventive decisions.

**Case studies:** The aim of case studies is to investigate the pattern of spreading a disease under specific conditions by examining a set of factors and individuals. In these studies, there is no allegation to extend the results to a bigger population.

**Control studies:** In control studies, the individuals of the study are grouped based on the specific diseases and they are compared with a controlled

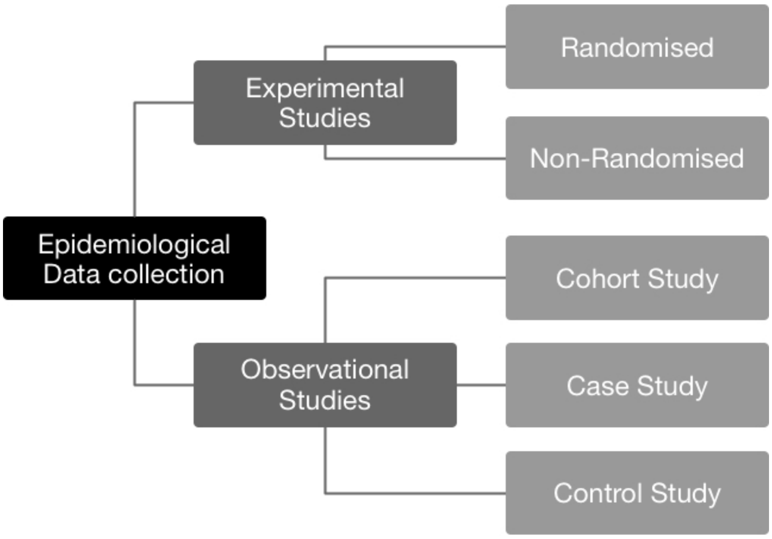


Figure 2.2: Taxonomy of epidemiological studies.

group of unexposed individuals, for example, a group who experienced an infection vs. healthy people.

### 2.1.1 Cohort Study Data Sources

The cohort study data are collected from various sources to describe individuals' characteristics in a population. The common sources of the data are as follows :

- Socio-economic and demographic findings: This information is mostly collected by questionnaires and interviews of the study participants. Socio-demographic refers to features like age, size, gender, marital status, and information like having pets. Socio-economic usually includes pieces of information on the level of education, monthly income, house type, and occupation pattern [8].
- Physical examinations: The physical characteristics of individuals that are associated with their health condition are examined or assessed and are saved to the database. Some physical characteristics consist of height, weight, BMI, blood pressure.
- Laboratory tests: The tests that may be associated with a disease are examined by blood or urine tests. These features are including glucose and blood sugar.

- Medical records: The medical history of individuals may give information about the current health condition of the participant. Thus, the data regarding experienced diseases, treatments as well as used medicaments are collected and saved in the database.

- Medical images: Sometimes the information from medical images like MRI or X-ray is extracted and annotated by an expert in a form to be able to be saved in the database. This information reflects the health status of an individual regarding a specific disease, e.g. fatty liver. However, collecting information from medical images is not trivial due to the costs and the time of the procedure.

### 2.1.2 Cross-Sectional Study vs. Longitudinal Study

**Cross-sectional:** In this type of study the data of a specific population and the pattern and relationship between features are collected in a single time point. In cross-sectional studies, the participants are not tracked for changes over time. In fact, the main task here is to analyze the participants in a single given time point. For example, analyzing the relationship between diabetes and smoking based on the comparison of sub-cohort $A$ aged between 20 and 30 and sub-cohort $B$ with participants between 31 and 40 years. Thus, the analyst should check the diabetes status for smokers and non-smokers for sub-cohorts $A$ and $B$.

Although cross-sectional studies have no additional costs and their analysis is not influenced by issues like dropouts, they are not sufficient to find out the cause of disease due to the reflection of a specific event in the past. This is mainly because cross-sectional studies only consider a single time point and ignore what happens in the past or future.

**Longitudinal study:** In longitudinal studies, the data is collected repeatedly from the same participants over time. The process of collecting data in longitudinal studies may take some years or decades. The time spent on data gathering depends on the type of information that needs to be collected.

The longitudinal studies are helpful because they allow the analyst to detect any changes in participants' characteristics. This makes it possible to follow the participants' characteristics as a sequence of events. An ex-

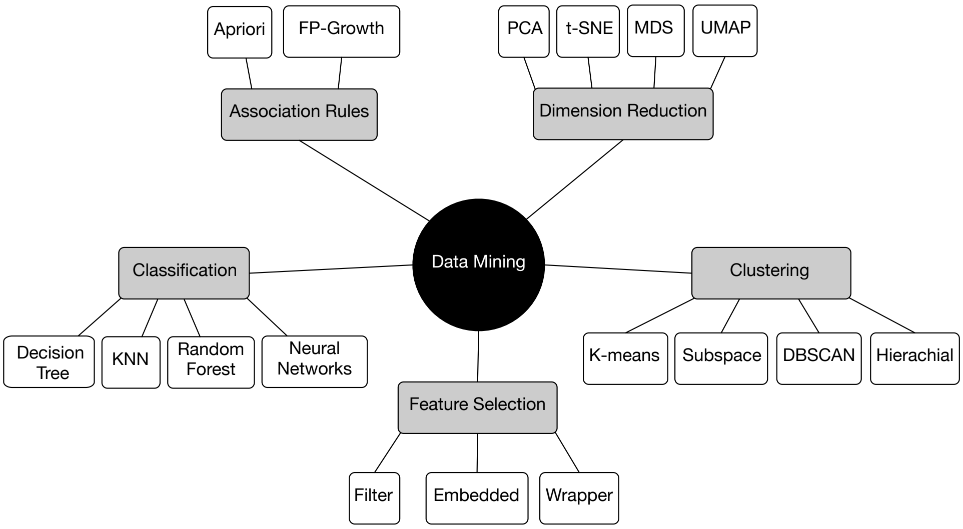


Figure 2.3: A general categorization of the most common data mining techniques.

ample of the usage of this kind of study is when the analyst wishes to observe the changes in the fatty liver status of women aged between 30 and 40 years who are smokers and follow these patients for the next 10 years. Therefore, it helps to understand the relationship between gender, smoking, and fatty liver.

Although from an economic point of view, longitudinal studies are more expensive than cross-sectional studies, they provide stronger results by involving detailed information of participants over time. However, cross-sectional studies are less challenging because there is no need for follow-up involvement of the same participants.

### 2.1.3 The Role of Data Mining in Epidemiological Studies

The data mining techniques as a part of computer science are applied in different research areas. They aim, to identify patterns in the data or prediction of an event by incorporating a different set of algorithms, statistics, and mathematics. One of the main goals of this thesis is to apply data mining techniques along with visualizations to analyze the data. In the following, some data mining techniques are explained, see Figure. 2.3.

### Classification

Classification is a supervised learning task to predict the class/label/target of a given data point. The classification consists of two steps: training and test. In the training step, the data (usually a part of the dataset) is used for the learning of the classifier; then, the test data assesses the accuracy of the classifier. In the following the well-known classification algorithms are described [9].

- **k-Nearest Neighbor (KNN):** The KNN is a common classification method where each observation is analyzed based on its closest neighbor and inherited its label from the nearest neighbor. The KNN technique is very intuitive and easy to implement, but it is not the best choice for the classification of high-dimensional data. We used a visual classifier in Chapter 4 for the discovery and validation of sub-cohorts based on the KNN algorithm. The overall steps to perform KNN classification are as follows:

  a For each sample in the database, calculate the distance between samples and keep the distances in an array.

  b Sort the array of each sample based on the distances in ascending order.

  c Pick the first K entries from the k selected entries.

  d Return the mode of the labels of the K selected entries.

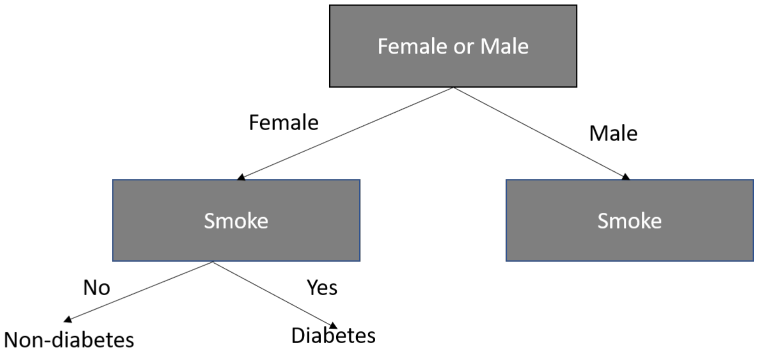


Figure 2.4: Example of a simple decision tree.

- **Decision trees:** The decision tree is a popular classification technique that classifies the data in a tree structure by applying if-then rules. The tree structure is built based on a top-down recursive approach. The decision tree is only applicable to categorized data, thus to perform it on continuous data, it should be discretized in a preprocessing step. Figure 2.4 shows a simple example of decision trees.

The main issue with decision trees is overfitting. It may lead to overgrowing the tree in many branches in the presence of outliers. There are some pruning techniques to manage this problem to prevent the tree from growing or pruning the unnecessary branches after growing the tree. Iterative Dichotomoiser 3 (ID3) and C4.5 are the most famous decision tree algorithms. ID3 proposed by Ross Quinlan in 1986 [10].

ID3 has a top-town greedy strategy and in each iteration, the algorithm selects the best features, i.e. divides the features into groups of two or more which is called a node. In general, ID3 construct a decision tree in the following steps [11]:

1. Get the dataset with $n$ features as input.
2. For each $i$ feature in the dataset.
3. It calculates the amount of certainty by entropy and information gain of feature $i$.
4. It selects the feature with minimum entropy and maximum gain as the best feature $best$.
5. Split the dataset based on feature $best$.
6. Create a node in the decision tree for $best$ feature.

ID3 has several drawbacks: first, the usability of ID3 is usually limited to only categorical features. Second, because of the overfitting it usually works well with the learned dataset, but fails to adapt and generalize with new datasets.

In 1993, Quinlan proposed C4.5 as an extension of ID3 to tackle the drawbacks of ID3. C4.5 is able to handle both continuous and categorical values, e.g. it uses a threshold value to discretize the continuous values. Moreover, it will prune the decision tree by removing branches that have less influence on the classification of the data.

Node-link/ Network diagrams and icicle plots are conventional visualizations of the decision trees. The node-link diagrams use a network representation to show the connection (by links) of the nodes (usually circles/dots/icons). Icicle plots are used to show hierarchy information; originally they are used to show the hierarchical clustering result [12].

BaobabView [1] is a famous tool for construction, pruning, and visualizations of decision trees [13]. Figure 2.5 shows a screenshot of BaobabView that uses node-link diagrams to represent the decision trees.

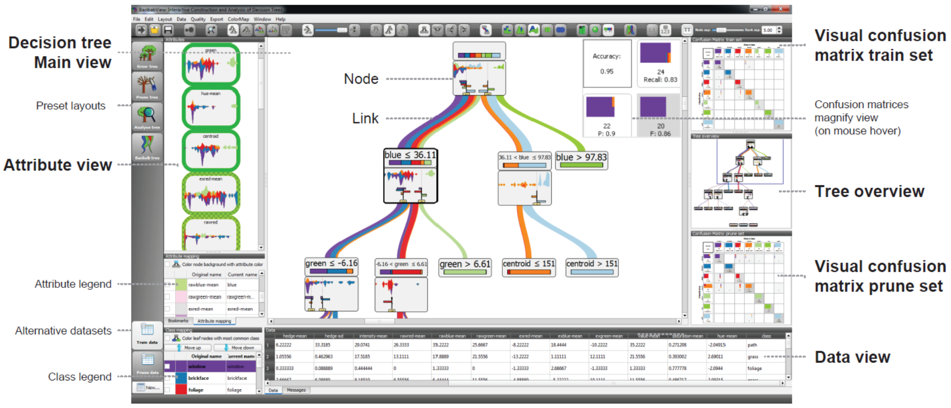


Figure 2.5: The interface of the tool for visualization of decision trees proposed by [13]

- **Random Forest:** Random forest is the enhanced version of the decision tree as it manages the problem of overfitting by selecting and merging the best decision trees. The random forest consists of a large number of decision trees, as each tree implies a class prediction. The model of prediction will be assessed by the class with the most votes. The key to build the result is the low correlation between the models. Thus, the random forest can be used as a solution for both classification and regression problems. Generally, the random forest algorithm has four main steps (see Fig. 2.6):

  1. In the first step, a portion of random samples is split from the given dataset as the training dataset.

---

[1] Baobab is an African short tree with an enormously thick trunk. It can live to a great age

2. Second, a decision tree is constructed for each sample of the training dataset. Then, the prediction results are assessed from every decision tree.
3. Next, the algorithm will make a voting for all predictions.
4. In the last step, the most voted prediction result will be selected as the result for the final prediction.

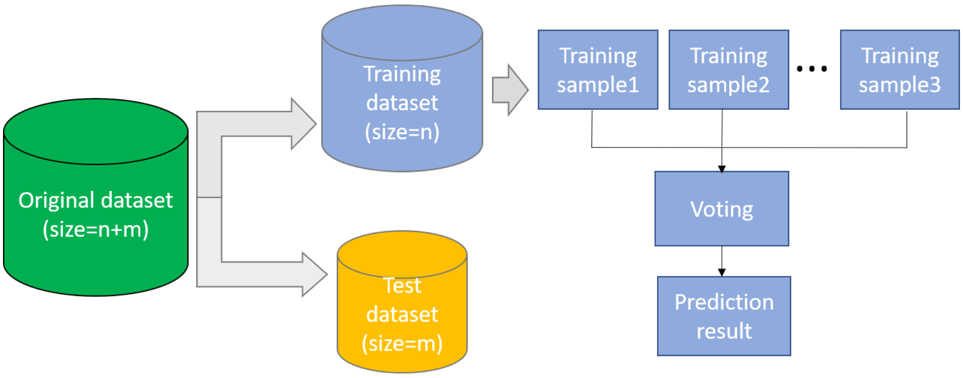


Figure 2.6: Random forest algorithm steps.

- **Neural networks:** The artificial neural networks are built based on neuron elements inspired by the human neuron system. They consist of multiple layers of neurons where each neuron takes a real value which is multiplied with a weight [14]. Although neural network classification works well with high-dimensional data and is able to find the complex relationships between features, it is non-trivial to be implemented as it needs a large dataset for training as well as high computational power for training the model. Generally training a network consists of the following steps:

  1. Split the input data to train and test subsets.
  2. The training data, i.e. data and an array of the corresponding labels for training, should be fed to the model.
  3. The model trains to find the relationship between the data and the labels.
  4. The test dataset should be fed to the model to predict the labels.
  5. Evaluate how the predictions match the labels of the test dataset.

### Feature Selection

- **Filter approach:** In the filter approach or single-factor analysis, the power of each feature for prediction is evaluated. The predictive power can be evaluated in various ways, such as getting the correlation between features and the target feature. The target feature is the one that should be predicted.
- **Wrapper approach:** The wrapper approach uses a combination of features to assess the predictive power. The most used wrapper approaches are forward step-wise, backward step-wise, and subset selection. The aim of the wrapper approach is to find the best combination of features. This technique is not suitable for large datasets as it is highly computationally intensive.
- **Embedded approach:** The embedded approach assigns weights to features. The features which are not good predictors get a low weight. In contrast, the features with a high predictive role get a higher weight. The regression techniques such as LASSO regression and RIDGE regression are used for the weighting of features.

### Association Rule Mining

The aim of association rules is to identify a set of items or features that occur together in a dataset. The term $itemset$ refers to a set of items together, i.e. consisting of more than one item. The frequent itemset mining technique identifies the itemsets that appear together.

Association rule methods discover interesting relationships between features in a dataset based on some interestingness of measurements. If $T$ is a set of transactions, $X$ and $Y$ are the itemsets and $X \Rightarrow Y$ are an association rule; then, the most important evaluation measurements to assess the interestingness the discovered rules are as follows [15]:

- Support: The $support$ factor indicates how frequently the items in a specific rule appeared together.

$$\mathrm{supp}(X) = \frac{|\{t \in T; X \subseteq t\}|}{|T|} \tag{2.1}$$

- Confidence: The $confidence$ factor represents the reliability of a specific rule. It indicates the percentage of the rule found true.

$$\mathrm{conf}(X \Rightarrow Y) = \mathrm{supp}(X \cup Y)/\mathrm{supp}(X) \tag{2.2}$$

- Lift: The $lift$ value reflects the importance of a specific rule. It is commonly computed by the equation below.

$$\mathrm{lift}(X \Rightarrow Y) = \frac{\mathrm{supp}(X \cup Y)}{\mathrm{supp}(X) \times \mathrm{supp}(Y)} \tag{2.3}$$

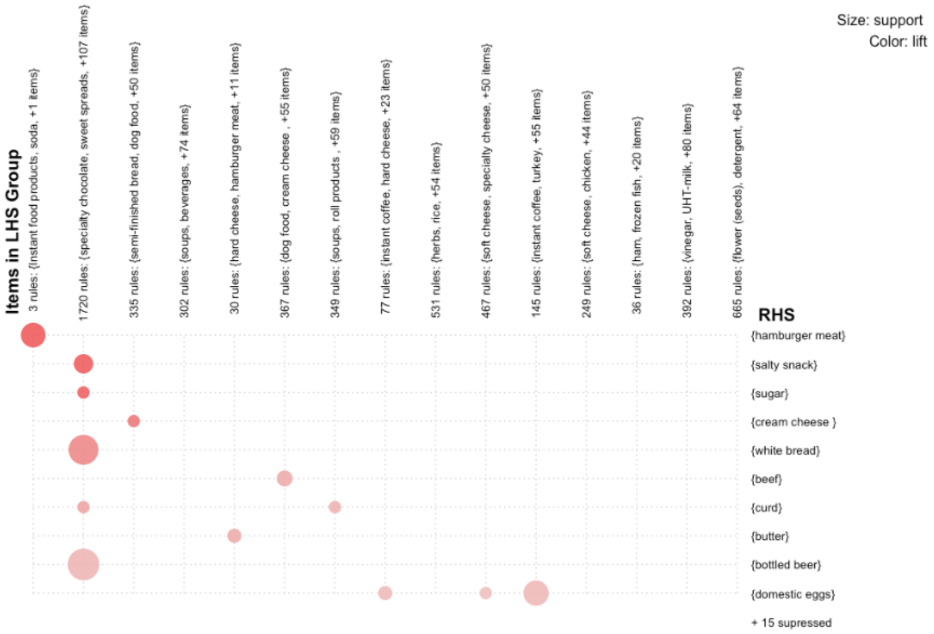


Figure 2.7: An example of the visualization of association rules using arulesViz package in R [16]

We used associate rules in Chapter 3 to find shared characteristics of participants who left the study.

- **Apriori algorithm:** The Apriori algorithm is one of the oldest and famous association rule mining algorithms proposed by Agrawal and Srikant in 1994 [17]. Apriori uses a bottom-up approach and iteratively finds the frequent data items in a dataset. The input of the algorithm is the dataset and a minimum support threshold which is set by the user. It identifies the itemsets by the following steps:

1. First, consider each item as a 1-itemsets candidate. Then, the occurrence of each item is counted by the algorithm.
2. Pick the candidates that satisfy the minimum support specified by the user for the next iteration.
3. Combine dual items and discover 2-itemset frequent items that satisfy minimum support.
4. Prune the 2-itemset candidate by minimum support value.
5. Continue the combining and pruning steps to find k-itemsets that satisfy minimum support.
6. By meeting the criteria, finding the most frequent itemsets, the algorithm will stop.

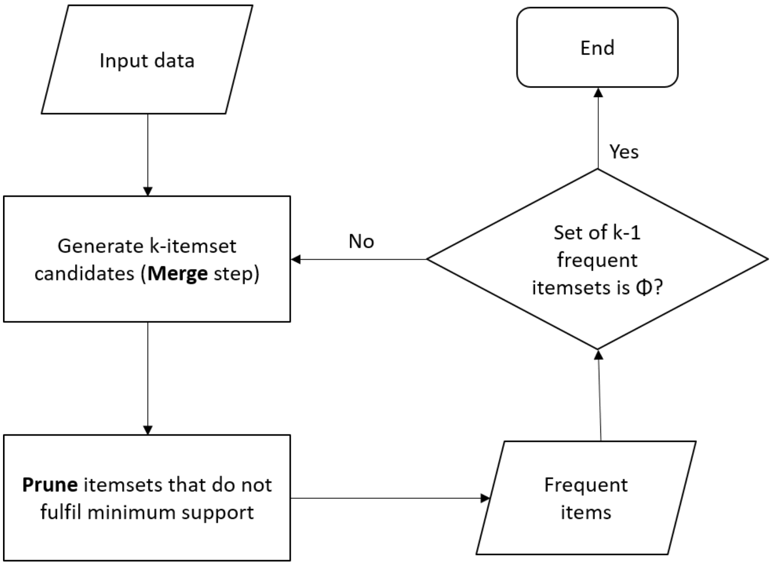

Input data
End
Yes
Generate k-itemset candidates (**Merge** step)
No
Set of k-1 frequent itemsets is Φ?
**Prune** itemsets that do not fulfil minimum support
Frequent items

Figure 2.8: The steps of the Apriori algorithm.

- **FP-Growth algorithm:** The FP-Growth is another popular and high-performance association rule mining algorithm proposed by [18] to find frequent patterns. The overall steps of the FP-Growth algorithm are as follows:
  1. Scan the database for the first time to find a frequent $1-itemset$, i.e. single item pattern.
  2. Sort the frequent items based on frequency in descending order, f-list.

3. Construct the FP-tree by re-scanning the database.
4. Sort the conditional FP-tree in the reverse order of the f-list to a generate frequent item set.

In Chapter 5, we use the FP-Growth algorithm to find out the set of patients who frequently appeared together in regions of interest of the DNA sequence.

**Dimension Reduction**

Dimension reduction techniques appeared in the early 20th century to tackle analytics of high-dimensional data. In general, it is obvious that dimension reduction techniques are applied to reduce the number of features of a given dataset without losing too much information.

Nowadays, we are generating a huge number of data in our daily life, as about 90% of the data around the world was only produced in the past four years. Some sources of the data include:

- Smartphones' apps collect a lot of personal information
- Google servers save the earth planet and searches
- Facebook collects all information that we share such as our clicks, likes and, etc.
- Instagram saves photos, videos, and our likes.

The healthcare and epidemiological datasets are no exception. They store health-related information of a population which leads to datasets with hundreds/thousands of features.

One may ask why we need to reduce the number of features. There are several reasons:

- By reducing the number of features, less space is required to store the data.
- For analysis, training a dataset with fewer features needs consequently less computation time.

- Some techniques, e.g. global clustering, do not work correctly with a high number of features.
- Most of the time in huge datasets some features are highly correlated, thus dimension reduction techniques solve the problem of collinearity by removing redundant features.
- It is easier to visualize low-dimensional data. Suppose that we have two features of age and height, it is really easy to plot the relationship between them in a scatter plot. In reality, we have many more features in a dataset, i.e. hundreds. Now, the challenge is how to visualize for example a dataset with $n$ features. It does not seem to be a good way to visualize pairwise relationships of features as we need $n(n-1)/2$ plots. It takes a lot of space and is difficult to follow.

Dimension reduction techniques are widely used in many fields, e.g. for the analysis of multi-omics data [19]. Figure 2.12 shows the usability of dimension reduction techniques for analysis of clustering results [20]. Analysis of high dimensional data is always challenging, because for the following reasons:

- A problem is the curse of dimensionality which means that with the increasing number of dimensions in a dataset the distance between points becomes vague. It means by adding more dimensions, the observations spread out from each other until the distance between them becomes equal. Figure 2.9 illustrates this problem. Thus, the clustering on lower dimensions of the data is more suitable for the analysis [21].

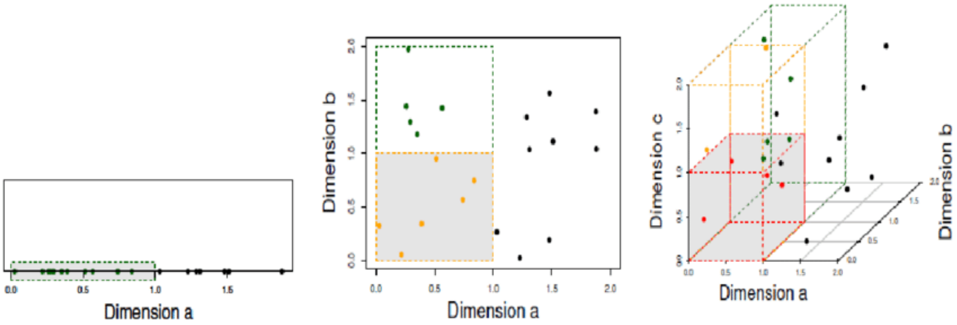


Figure 2.9: By adding more dimensions, the sparsity increases exponentially [21].

- With large datasets the training and analysis of the data is slow.
- A large dataset needs a bigger storage space.
- Usually, the data might be noisy and with a lot of irrelevant and redundant features.

Thus, dimension reduction techniques help to address all the above-mentioned issues by reducing the feature space. Some dimension reduction techniques are described subsequently.

- **PCA:** The principal component analysis is a linear well-used dimensionality reduction technique [22]. This technique generates a set of new features from a given huge dataset with a high number of features. The new set of generated features is called the principal component. The most common characteristics of PCA are that:

  a The principal components are linear combinations of the original features.

  b The first principal component represents the maximum variance in the dataset and the second principal component's aim is to describe the rest of the variance in the dataset which is uncorrelated to the first principal component.

  c The next principal component describes the variance of the dataset which is not included by the previous principal components.

  In Chapter 5 the PCA is used for 2D projection of mRNA data and to visualize the similarity between sub-cohorts. In PCA the high variance features form the principal components and the data will be rotated along the direction of higher variance.

  In fact, each principal component is a linear combination of the original predictors of the given dataset. As an example, suppose that we are having a dataset with a set of $p$ predictors and we want to extract the two principal components $PCA1$ and $PCA2$.
  If a set set of predictors is:

$$Normalized\ Predictors = X^1, X^2, ..., X^p \tag{2.4}$$

The first principal component is calculated as:

$$PCA^1 = \Phi^{11}X^1 + \Phi^{21}X^2 + \Phi^{31}X^3 + ... + \Phi^{p1}X^p \tag{2.5}$$

where
$\phi p^1 = (\phi^1, \phi^2, \phi^3, ...)$ is the loading vector for the first principal component. As the large value of loading may lead to a large variance, the sum of the squares of the loading vectors is limited and is equal to one. The result of the $PCA1$ in the $p - dimensional$ space is a line that is the closest to the n observations based on the (usually) Euclidean distance. In other words, the resulted line minimize the distance between the line and a data subject.

The second principal component $PCA2$ can be calculated the same as $PCA1$, see Eq. 2.6. The $PCA2$ is uncorrelated with $PCA1$ as their correlation is equal to zero and it explains the remaining variance of the dataset.

$$PCA^2 = \Phi^{12}X^1 + \Phi^{22}X^2 + \Phi^{32}X^3 + ... + \Phi^{p2}X^p \tag{2.6}$$

When we visualize the two uncorrelated components, they have orthogonal directions to each other. Figure 2.10 shows the PCA calculation of RNA data in Chapter 5.

- **MDS:** Multidimensional scaling is also a popular linear dimension reduction technique [23]. This technique computes the similarity of the data points and projects the points in lower feature space based on a distance function, i.e. Euclidean distance, as closer points in high-dimensional data are also closer to each other in lower dimensions. MDS is applied in Chapters 4 and 5 to show the similarity of sub-cohorts. Following are the steps of MDS:
  - a Set the number of dimensions after reduction in n-dimensional space. In n-dimensional space, this number could be between 2 and 3 (dimensions more than 3 are difficult to visualize). Then, set the orientation of the coordinates north, south, east, and west.
  - b The distance of all pairs of points in the dataset should be calculated using Euclidean distance. The result is an $n * n$ similarity matrix, where n is the number of observations in the dataset.

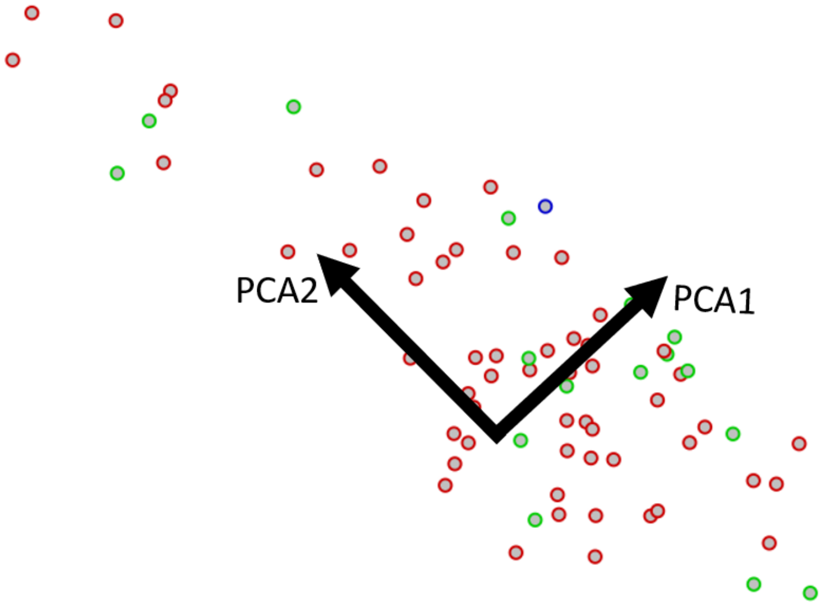


Figure 2.10: PCA shows uncorrelated components having orthogonal directions to each other.

c Evaluate the stress function by comparison of the similarity matrix and original dataset, where stress is a measure for goodness of fit according to the predicted and actual distances.

d Adjust the coordinates to minimize stress.

- **t-SNE:** PCA and MDS are linear approaches, but what if we want to find a pattern in a non-linear approach? T-distributed Stochastic Neighbor Embedding is a non-linear dimension reduction technique that calculates the probability of the relationship between data points in high-dimensional data, then mapping the points with a similar distribution in the low-dimensional space [24].

  Mainly there are two methods to map the data points in low-dimensional space, local and global. T-SNE maintains the local and global structure simultaneously. Local techniques map each data point by nearby data points in low-dimensional space. The second method is a global technique that preserves the mapping of points on the manifold of near points and keeps the distance of faraway points to faraway data points. In other words, it aims to keep the geometry at all scales, including low-dimensional space and high-dimensional space. It converts the distance between data points in the form of probabilities.

In general, the t-SNE calculates the distance between data points in three steps:

1. To calculate distances in high-dimensional space, t-SNE calculates the Euclidean distance between data points in the form of probabilities which are representative for similarities of data points. The distance probability between point $x_i$ and $x_j$ is calculated according to 2.7, where $\sigma$ is the variance of high-dimensional data .

$$p_{j|i} = \frac{exp(-||x_i - x_j||^2/\sigma_i^2)}{\sum_{k \neq i} exp(-||x_i - x_k||^2/\sigma_i^2))} \tag{2.7}$$

2. In the low-dimensional space, the distance probability between data points $y_i$ and $y_j$ (corresponds to $x_i and x_j$ in high dimensional space) is calculated according to Equation 2.8.

$$q_{j|i} = \frac{exp(-||y_i - y_j||^2)}{\sum_{k \neq i} exp(-||y_i - y_k||^2))} \tag{2.8}$$

3. Minimizing the difference between probabilities in high-dimensional space and low-dimensional space

In Chapter 5, the t-SNE is used as an option for 2D projection of RNA data.

- **UMAP:** Although t-SNE works well with mid-size datasets, it has its own limitations like computation time in the case of having very large datasets.

  The Uniform Manifold Approximation and Projection (UMAP) is a recent dimension reduction technique [25] which is an enhanced version of t-SNE. It managed to solve the issues with t-SNE. UMAP works well with very large datasets and preserves both the local and global structure of the data points. UMAP works with uniformly distributed data based on Riemannian geometry. The distance between data points can be calculated by:

1. Calculating the distance between the data points in high-dimensional space.
2. Calculating the distance between the data points in low-dimensional space.
3. Minimizing the difference between both distances using Gradient descent.

We used the UMAP algorithm as an option for projection of mRNA data in Chapter 5.

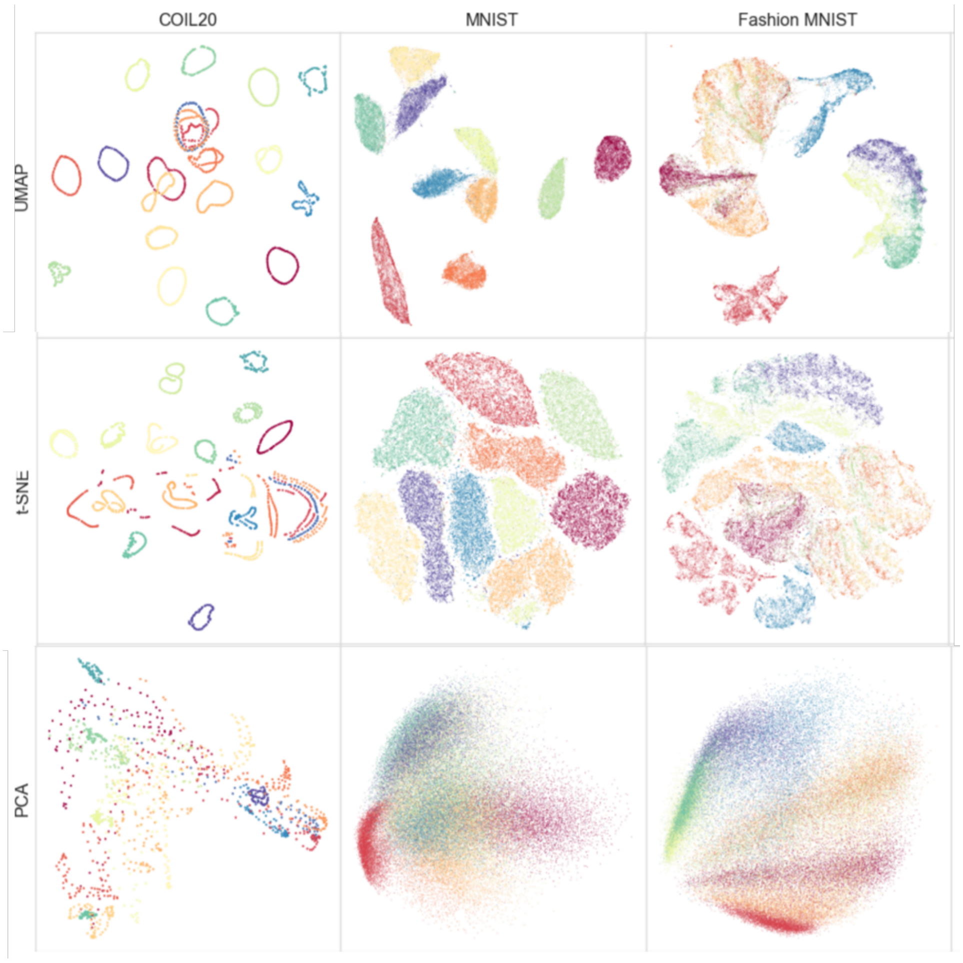


Figure 2.11: Comparison of different dimension reduction techniques on different datasets [25].

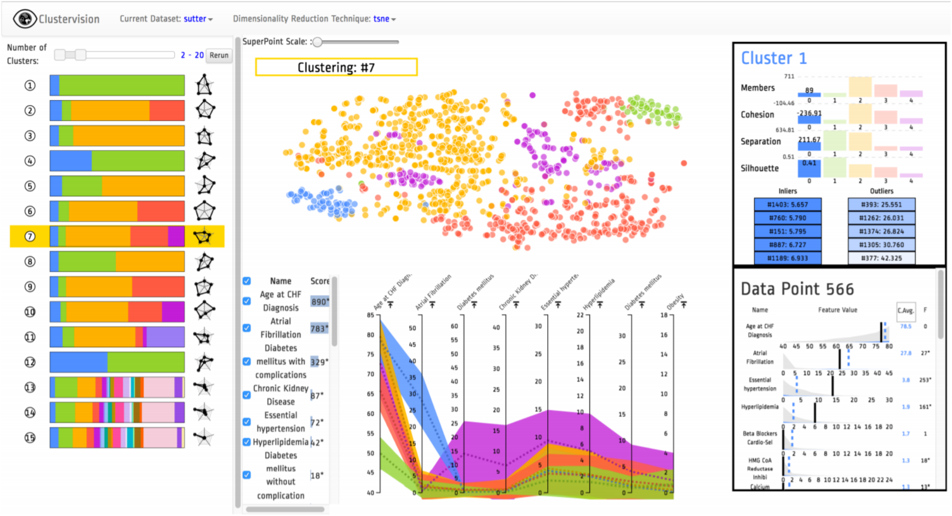


Figure 2.12: A screenshot of ClusterVision for analysis of clustering results [20].

### Clustering

Clustering is one of the most important techniques in machine learning. It is an unsupervised technique with the aim of identification of the correlation between features in the data with minimum detailed information about the data. The outcome of clustering can be used to generate hypotheses. The most important clustering algorithms are:

- **K-means:** The K-means clustering partitions the data into k clusters, as the observations with the nearest mean, belong to the same cluster [26]. In fact, it aims to minimize the variance of observations within a cluster. Suppose having n observations $\{(x_1, x_2, x_3, ..., x_n)\}$ and d features, the k-means algorithm will group the n observations into the $k$ sets $C = \{C_1, C_2, ..., C_k\}$. Thus, the main goal is to find:

$$arg_c \min \sum_{i=1}^{k} \sum_{x \in C_i} \left\| X - \mu_i \right\|^2 = arg_c \min \sum_{i=1}^{k} |C_i| \, Var C_i \qquad (2.9)$$

  where $\mu_i$ is the mean of points in $C_i$. The K-means clustering is used in Chapetr 3 in a step for generating the dummy dataset with missing data.

- **DBSCAN:** The density-based spatial clustering of applications with noise (DBSCAN) is a widely used density-bound clustering technique [27] and [28]. The density-based techniques identify the areas with high density as clusters. The density of data points within a cluster is higher than outside of the corresponding cluster.

To measure the density of a specific data point, DBSCAN takes two main parameters to be applied in clustering: $eps$ specify how close the observations should be in a dense region-based using a specific distance measure, and $minPoints$ determines the minimum number of connected observations in a region. DBSCAN has the strength to identify outliers.

It is important to set the parameters to an appropriate value. If we have a dataset with $n$ data points, by setting $eps$ to a very large value, all data points will have a density $n$. The reason is that for each data point all other points will lie in the $eps$ area of the given data point. In contrast, if we set $eps$ to a very small number, all data points will have density 1.

The data points in a DBSCAN approach can be in one of the three following groups, see Fig. 2.13:

- **Core point:** Core point is a data point that always belongs to the dense region. The number of points around a key point is more than the defined number for $minPoints$ within the $esp$ region.
- **Border point:** If a point is in the neighborhood of a core point, but has fewer points than $minPoints$ within the $esp$ area, it is called a border point.
- **Noise point:** Any point which is neither a key point nor a border point, is called a noise point.

The key points whose distance is maximum $eps$ lie in the same cluster, while border points will be assigned to the cluster of its neighbor key point. The noise points will be ignored and considered as outliers. Overall, DBSCAN works as follows:

a Identify and label key, edge, and noise points.

b Exclude noise points.

c Connect the core points which lie in a sphere within $eps$ area.

d The connected core points will form a cluster.

e Assign each edge data point to its neighborhood key point.

The DRESS algorithm that we use for the subspace clustering of epidemiological data in Chapter 4 is a DBSCAN-based algorithm.

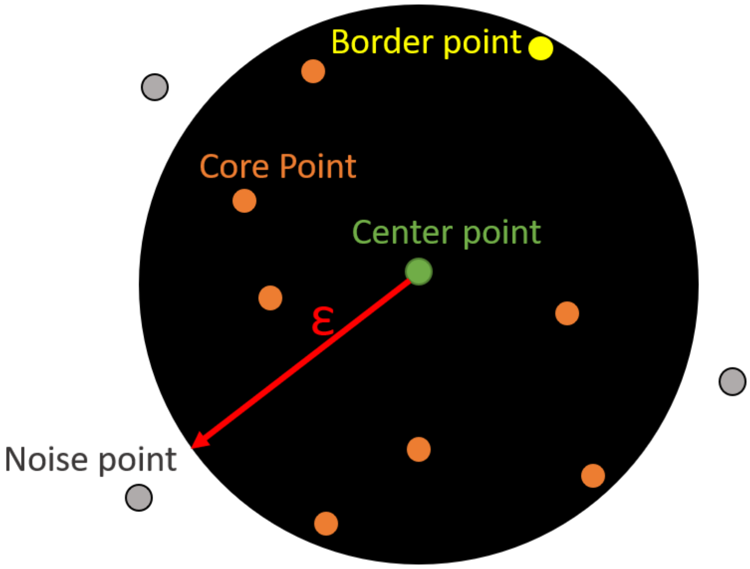


Figure 2.13: The circle shows the $esp$ area of a DBSCAN algorithm.

- **Hierarchical:**

  Hierarchical cluster analysis groups similar observations into clusters. The result of hierarchical clustering is a set of distinct clusters where the observations within each cluster are similar to each other.

  Hierarchical clustering builds a hierarchy cluster based on a specific strategy, either agglomerative or divisive [29]. $Agglomerative$ is a bottom-up technique, which means that clustering starts with each observation (each observation forms a cluster), and as we move to an upper level of the hierarchy it merges the pairs of clusters. $Divisive$ techniques are called top-down, where it starts with one cluster of all observations and in each move down the iteration the cluster will split up. The result of both techniques is visualized in a dendrogram. The hierarchical clustering is used for the clustering of RNA samples of cancer patients in Chapter 5.

- **Subspace:** Subspace clustering is a technique that considers all features in the clustering process, and the algorithm finds similar ob-

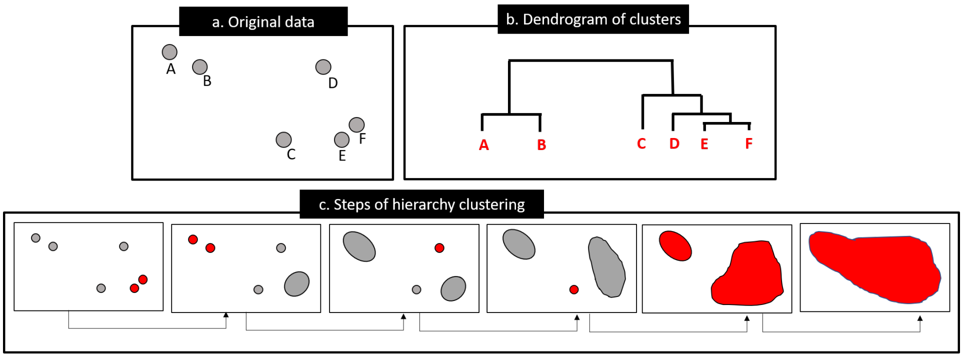


Figure 2.14: a) Shows the data points before clustering. b) Dendrogram of clusters after hierarchical clustering. c) The iterative steps for hierarchy clustering of data points in (a).

servations that can be grouped together in an overlapped subset of dimensions [21]. There are a number of works that review the (subspace-)clustering approaches. In fact, subspace clustering is the enhanced version of the traditional clustering technique. Its main aim is to find groups of objects in different subsets of dimensions, i.e. subspaces, without eliminating some dimensions or objects. Overall, it has two main steps: the first is the feature selection and the second step is to attempt the clustering technique. In other words, subspace clustering needs a search method to evaluate each combination of subspaces to apply the clustering. A detailed description of subspace clustering and its application to find sub-cohorts in epidemiological data is explained in Chapter 4.

- **Biclustering:** Biclustering is a powerful data mining technique that emerged to allow simultaneous clustering of rows and columns. Biclustering techniques mainly were applied for clustering of omics data, however, it is believed that with gaining knowledge on choosing appropriate biclustering tools and enhancing the technique to be applicable on a wide range of the data in can be applied on other variety of the data [30]. Biclustering is considered a pattern-based technique, it means depending on the problem the relevant algorithm should be selected. In a generalized categorization, the algorithms are defined as [31]:

a **Bicluster with constant values:** The aim of this type is to find biclusters with constant values on rows and columns. In practice, this type of algorithm is not really applicable since, in reality, the data is often noisy. Hartigan algorithm is an example to find biclusters with constant values. Fig 2.15 (a) shows an example of constant values bicluster. A perfect constant values bicluster is a submatrix with size $(I, J)$, where for all $i \epsilon I$ and $j \epsilon J$ the values are equal and the variance is zero.

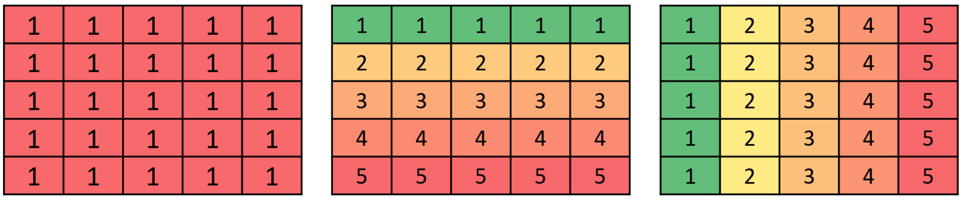
b **Biclusters with constant values on rows or columns:** These types of algorithms look for clusters with constant values on rows or columns. To apply this type of algorithms the data should be normalized in a preprocessing step. Many algorithms focus on this type of biclustering. Fig. 2.15 (b) and (c) shows examples of constant values on rows and columns in the best case. These types of biclusters can be obtained by $a_{ij} = \mu + a_i$ or $a_{ij} = \mu * a_i$ expressions, where $\mu$ is a constant value within a specific bicluster and $a_i$ is the adjusted value on $i$th row that can be achieved either by an additive or multiplicative way.

c **Biclusters with coherent values:** These kinds of biclusters are an improvement on the previous types which aim to find biclusters with constant values on both rows and columns (see Fig. 2.15(d-e). More advanced techniques are used to assess the quality of biclusters by calculation of covariance between rows and columns. These biclusters are also achieved based on two models, additive and multiplicative.

In Chapter 4, it is suggested to use biclustering for analysis of longitudinal epidemiological data as future work.

## 2.2 Study Examples

The most famous study examples in Europe are the Study of Heath in Pomerania and the Rotterdam Study.



| | | | | |
|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 |
| 1 | 1 | 1 | 1 | 1 |
| 1 | 1 | 1 | 1 | 1 |
| 1 | 1 | 1 | 1 | 1 |
| 1 | 1 | 1 | 1 | 1 |

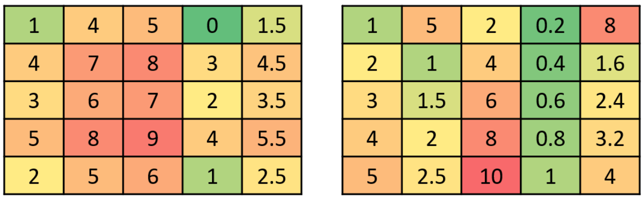
(a) Constant values

| | | | | |
|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 |
| 2 | 2 | 2 | 2 | 2 |
| 3 | 3 | 3 | 3 | 3 |
| 4 | 4 | 4 | 4 | 4 |
| 5 | 5 | 5 | 5 | 5 |

(b) Constant values on rows

| | | | | |
|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 |
| 1 | 2 | 3 | 4 | 5 |
| 1 | 2 | 3 | 4 | 5 |
| 1 | 2 | 3 | 4 | 5 |
| 1 | 2 | 3 | 4 | 5 |

(c) Constant values on columns



| | | | | |
|---|---|---|---|---|
| 1 | 4 | 5 | 0 | 1.5 |
| 4 | 7 | 8 | 3 | 4.5 |
| 3 | 6 | 7 | 2 | 3.5 |
| 5 | 8 | 9 | 4 | 5.5 |
| 2 | 5 | 6 | 1 | 2.5 |

(d) Coherent additive values

| | | | | |
|---|---|---|---|---|
| 1 | 5 | 2 | 0.2 | 8 |
| 2 | 1 | 4 | 0.4 | 1.6 |
| 3 | 1.5 | 6 | 0.6 | 2.4 |
| 4 | 2 | 8 | 0.8 | 3.2 |
| 5 | 2.5 | 10 | 1 | 4 |

(e) Coherent multiplicative values

Figure 2.15: a) Constant values bicluster. b) constant values on rows bicluster. c) Constant values on columns. d) Coherent additive values bicluster. e) Coherent multiplicative values bicluster.

### Rotterdam Study

The Rotterdam study was appointed in the Ommoord region of Rotterdam city in the Netherlands in the late 1980s [32]. The main focus of this study was to analyze cardiovascular, endocrine, hepatic, neurological, ophthalmic, psychiatric, dermatological, otolaryngologic, locomotor, and respiratory diseases as well as diseases that are common in elderly persons, like Parkinson and Alzheimer disease. Generally, the Rotterdam study consists of four independent cohorts, and the data were collected in parallel. The first cohort (R-I) was established between 1990 and 2011 with 7983 participants. The second cohort study (R-II) was conducted from 2011 to 2016 with 3011 participants. The third cohort study (R-III) took place between 2006 and 2014 with 3932 participants. The fourth cohort study (R-IV) was established in 2016 and is still ongoing.

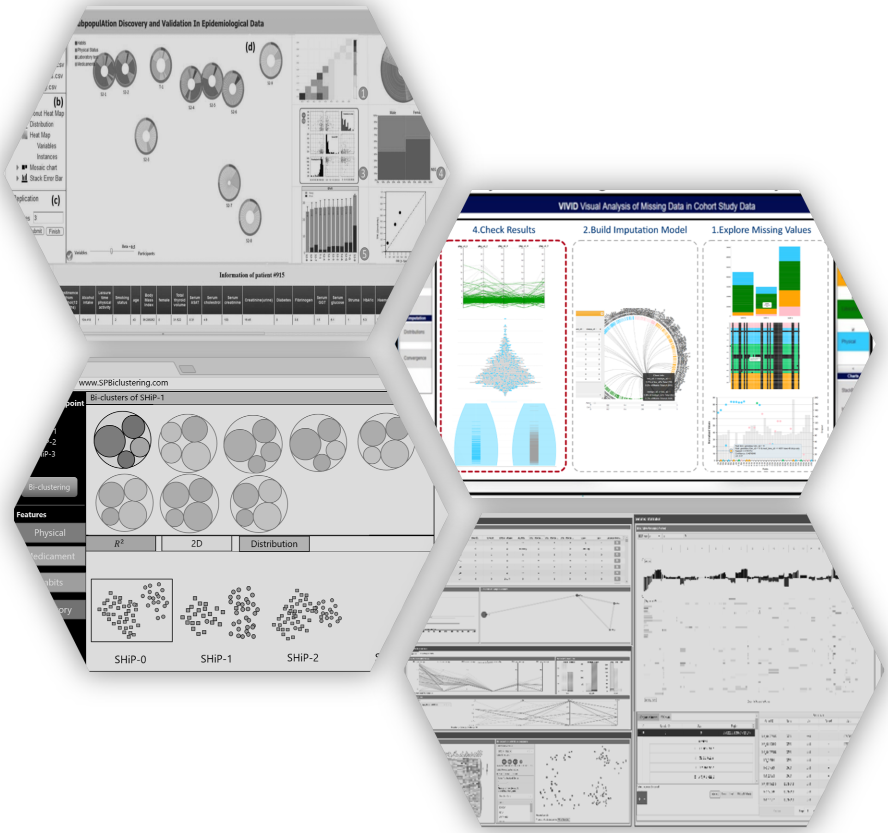
### Study of Health in Pomerania

The Study of Health in Pomerania (SHiP) is performed in the Northeast region of Germany, i.e. Greifswald, Stralsund, and Anklam [33]. The same as for other sources of epidemiological data, the information is acquired via interviews of participants and gaining information on sociodemographics, lifestyle, and medicaments. Moreover, some information is collected based on physical examinations, for example, BMI and blood pressure status, while some information is based on laboratory examinations, e.g. diabetes status and information on liver functionalities. Additionally, some information is extracted from medical images.

The SHiP study is carried out in Western Pomerania. The whole study consists of two dependent datasets: SHiP and SHiP-TREND. The main aims of the study were frequent diseases like fatty liver, breast cancer, and back pain.
The first collection of the first cohort (SHiP) started from 1997 to 2001, i.e. SHiP-0 with 4308 participants. The second wave of SHiP-1 was conducted from 2002 to 2006 having 3300 participants, and the third wave was between 2008 and 2012 with 2500 participants. The last wave of the study, i.e. SHiP-3, was performed between 2008 and 2012 with 1700 participants. As explained in the previous chapter, the study was initially started with 4308 participants, but within each next wave of the study, the number of participants reduced dramatically. Thus, a new study started in the same region with 8016 new participants, called SHiP-TREND.

# 3

# Data Quality



A fundamental topic in analytics is data quality, as it should be considered to have valid and accurate results. Thus, data cleaning is a vital preprocessing step to have valid outcomes [34]. As shown in Fig. 3.1, based on the taxonomy of dirty data/ rogue data issues by Kim et al. [35] data quality is a broad topic. Dirty data refers to inaccurate or incomplete data stored in the database. The inaccuracy of data can be due to misspelling or punctuation errors.

In many fields, during and after the collection process of the data, more specifically time-oriented data, the quality of data is questioned to ensure the accuracy of the analysis. One unavoidable aspect of the data quality is missing data, where some information for any reason remains unfilled. Consequently, it likely leads to biased data, since the available data is not representative for all real data and it restricts the statistical power of further analysis. As in many research fields, this problem is always present in epidemiological data. There is a number of works to handle missing data, but still, there is a gap to reach a satisfactory level of managing the missingness issue, which may be filled by visual analytic approaches. Following, the related works for handling missing data and the role of the visualization in this topic are explained, followed by a more detailed description of missing values in epidemiological data.

## 3.1 Missing Values as a Data quality Issue

Missing data is considered as a data quality factor, since it may affect the results of the analysis. In longitudinal epidemiological studies, it happens within and between the waves of the study. The missingness happens within waves of the study when some information is not filled for any reason. This issue is called an item non-response. The missing data is present in longitudinal studies when the participants are re-invited for follow-up examinations but do not show up. The terminology of unit non-response or dropout is used for this case of missingness, which means that all information of the corresponding participant becomes unavailable from a certain point. There can be many reasons for each form of missing data. Since in the presence of missing data the dataset could not be represented as real and whole data, it leads to a number of problems. Depending on

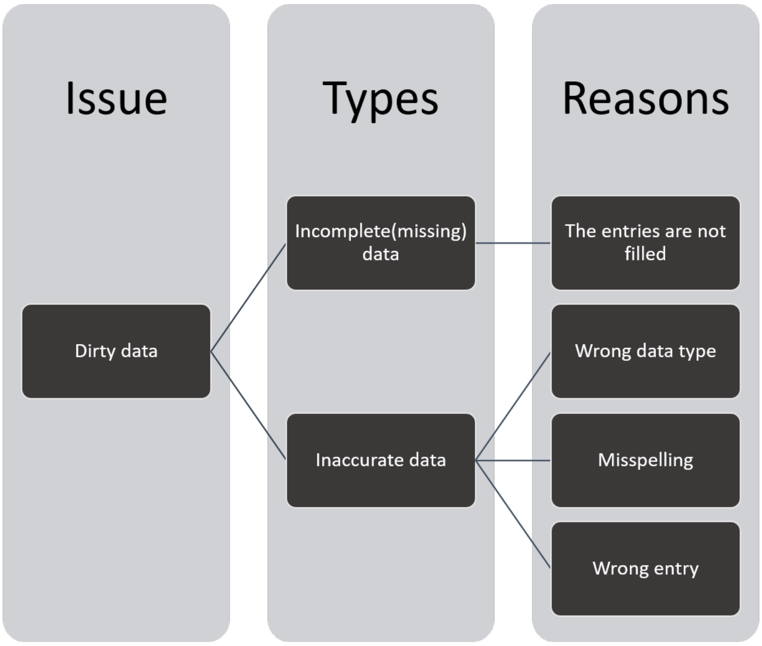


Figure 3.1: Taxonomy of data quality issues.

the number and type of missing data it reduces the statistical power of the study and leads to biased estimations [36].

There are various reasons for missing values within waves of the study. The way we handle missing values is determined by the cause of missing values because it explains whether the presence of missing values influences the distribution of the original dataset or not. The probable reasons for missingness within waves of the study are:

- Laboratory issues, e.g. blood's sample tube breaks accidentally
- Lack of biological samples, e.g. participant's fear to give a blood sample
- Errors in data entry, e.g. answers to some questions are ignored
- Non-response to certain questions/examinations, e.g. the participant denies answering specific questions for any reason

The possible reasons for missing data between waves of study include:

- Physically unable, e.g. disabilities to show up for examinations

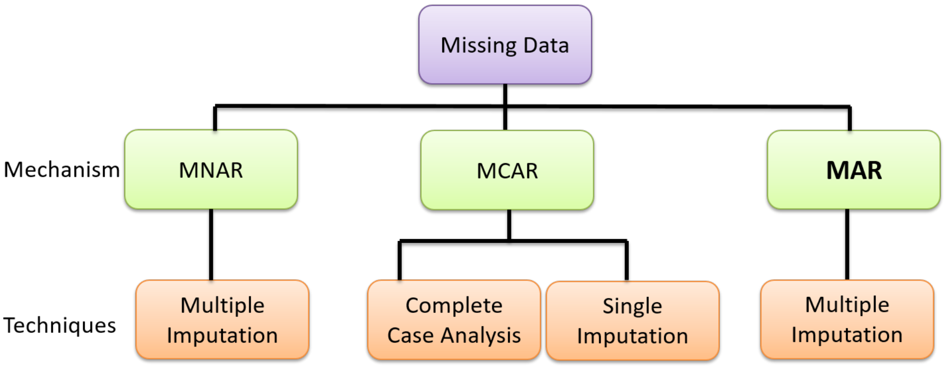


Figure 3.2: Taxonomy of the missing data mechanism.

- Inconvenient location, e.g. participants moved out from their former location
- Schedule conflicts, e.g. the determining date for examinations is not appropriate for participants
- Forgetting visits, e.g. not all participants recorded the follow-up examination date
- Side effects, e.g. the participants are not in a good health condition because of the side effects of some treatments

In general, there are three types of missing values and the way we should handle missing data depends on this missingness mechanism [37]:

- Missing Completely At Random (MCAR): when there is no logical explanation for the missing values. For example, if a participant accidentally skips a question in the questionnaire or the sample tube breaks, and consequently the result cannot be provided. In this case of missingness, usually, the distribution of missing values and observed values is the same and it does not lead to biased data.
- Missing At Random (MAR): this type of missingness occurs when the missing values are dependent on some available data. This type of missing value usually is a source of biased data. As a simple example, in a dataset where both features of age and blood pressure are considered, it is more likely that the blood pressure measure remains

unfilled for younger participants. The reason might be that it is less likely that younger people have hypertension issues. Thus, it might be ignored to be measured for young people. As a consequence, the observed blood pressure values tend to be higher (for elderly people) and missing for younger people with lower blood pressures.

- Missing Not At Random (MNAR): this type of missing data also leads to biased data as the distribution of observed and missing data is different. Like the MAR type, the missingness is dependent on what is happening in the study, but the cause of dependency is missing in the data (mostly to the specific feature itself). For example, in a study of migraine patients who are really sick and do not show up. As a result, information about the migraine status of people with headaches remains missing.

## 3.2 Related Works in VA

In this section, a brief explanation of previous works on handling missing data in epidemiological studies is provided. Then, the focus is on visualizations related to works with data quality topics.

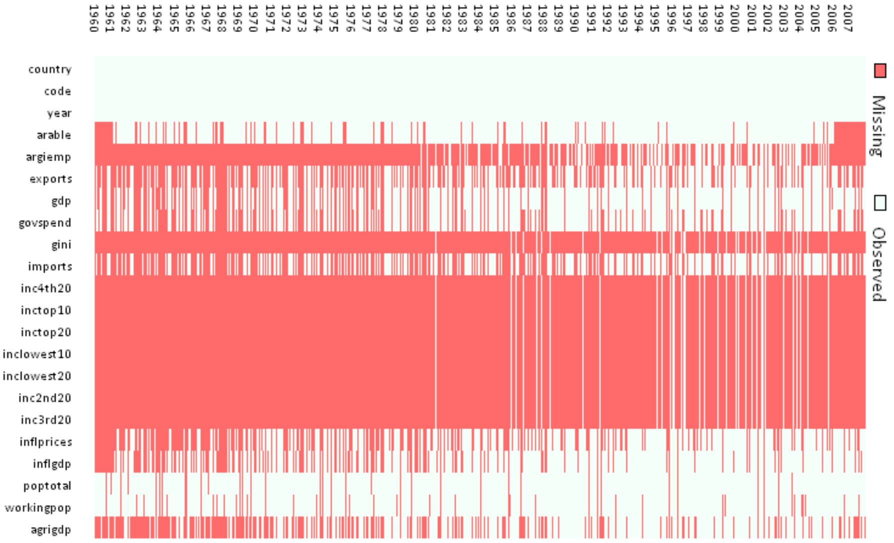


Figure 3.3: Missingness map of Amelia package [38]

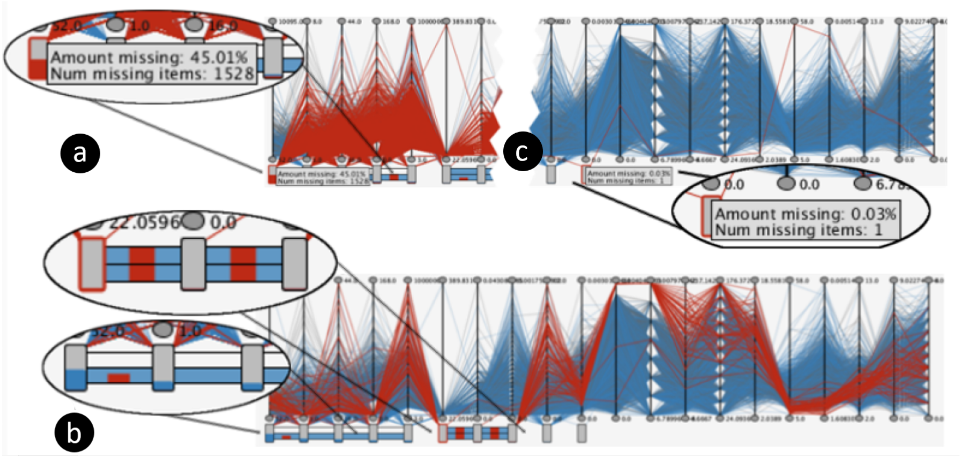


Figure 3.4: Missingness pattern in parallel coordinates (the red highlights represents the missing values) [39].

### 3.2.1 Handling Missing Epidemiological Data

There are various methods for handling missing data. However, depending on the type of missingness, the handling method should be selected carefully to not produce bias to the data [37]. The distinction between different types of missing data is not easily achievable, but the exploration of the missing data can help to get a sense. Usually, the distinction between MAR and MNAR is based on the assumption and background knowledge of the healthcare specialist. Figure 3.2 shows the taxonomy of different approaches for handling missing data depending on the missingness mechanism. The description of each technique is listed below:

- Single imputation: This is one of the simplest and common approaches to replace the missing data with a single predicted value. There are multiple single imputation methods:
    - Mean imputation, where all the missing values of a feature are replaced with the mean value of the corresponding feature. This imputation technique works well when the number of missing values is small and also the distribution of missing values and observed values is the same (MCAR).
    - Complete Case Analysis (CCA): In this technique, missing the subjects with missing data in one or more features is ignored

  from the analysis. Although CCA is also widely used in many research areas to polish the data, it has the same consequences as mean imputation. Moreover, we loose the whole information of subjects who have missing data and it reduces the statistical power of the data. This case is only appropriate in the MAR case.

  The above-mentioned techniques are easy to use, but lead to a significant bias in data when the distribution of the missing data is different compared to the distribution of observed data, i.e. in the case of MAR and MNAR. There are other single imputation methods that consider the relationship between features (regression methods) in the prediction process. Since the main focus in this thesis is multiple imputation, the detailed description of such techniques is skipped.

- Multiple imputation: When there is a dependency for missing data (MAR and MNAR) and thus the distribution of missing data and observed data is different, multiple imputation seems to be a reasonable approach to handle the missing data [37]. The relationship between features should be considered to predict missing values by using regression methods. For each missing value, multiple data is predicted. The main idea of multiple imputation is that since the predicted value could not be representative for the real value, for each missing value multiple values are predicted. The difference between predictions represents the amount of uncertainty of predictions. Generally, as shown in Fig. 3.5, multiple imputation consists of three main steps [40]:

  In the first step, *m* imputed datasets are created using multiple imputation by replacing missing values with plausible values. The plausibility of predictions is assessed from the distribution of predicted values for missing entries. In all these *m* datasets the observed values are the same, while they are different for missing entries. In the second step, the plausibility of the imputed datasets should be checked through analytic methods like analyzing the difference between observed data vs imputed data. The last step is to pool the datasets into one, as it reflects an unbiased dataset and the imputed dataset simulates the distribution of missing data besides the observed data.

In this work, multiple imputation of chained equations (MICE) as a type of multiple imputation is used [41]. The following explains the steps of MICE:

- Applying a simple single imputation method to fill the missing values as place holders, e.g. mean value imputation.
- For all features $v$ a regression model should be learned for missing values on both imputed and observed values.
- Replace the missing values of $v$ with the predicted values from the regression model. These two last steps are repeated $maxit$ numbers to generate $m$ imputed datasets.

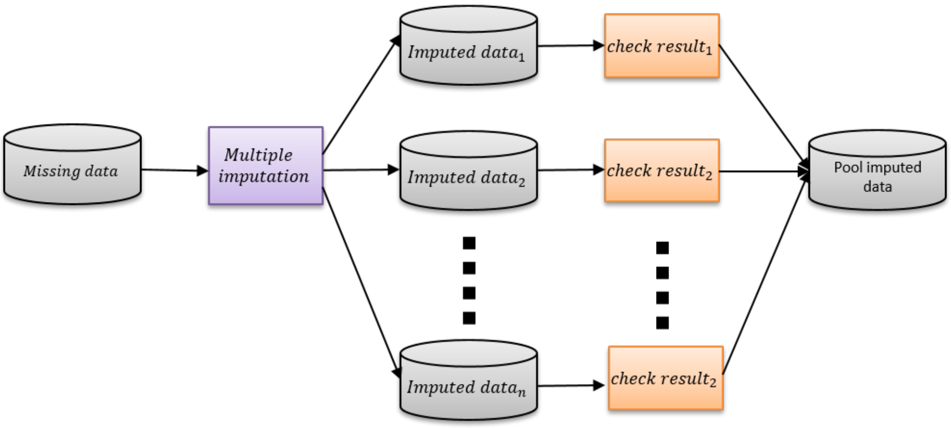


Figure 3.5: Multiple imputation main steps.

### 3.2.2 Task Analysis

As the main aim of this chapter is to handle missing values in cohort studies, the three main tasks that should be considered in treating missing values (exploration, imputation and check the results) are explained as follows [42]:

T1 What has to be answered in the exploration phase:

a. Which portion of the data is missing within and between the waves of the study?

  b. Where are the item non-response and unit non-response cases?
  c. Does missingness happen with the same pattern in a pair of features, i.e. co-missingness?
  d. What are the shared features of participants who dropped out of the study?

T2 When building the imputation model, which parameters should be considered?

  a. Can we make a trade-off between accuracy and computation time of imputations?

T3 Are the imputation results valid?

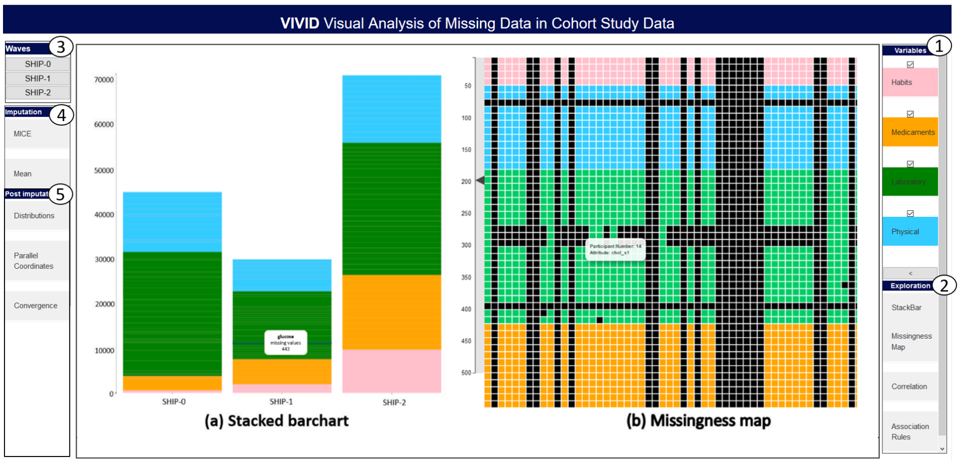
  a. Do the predictions for each feature make sense?
  b. How are the imputed values distributed over observed values considering min, max, dense and sparse region?
  c. How do the predictions differ in each $m$ imputed dataset for a selected feature?

### 3.2.3 VIVID

VIVID was developed as a web-based framework to fulfill the tasks described in Section 3.2.2. The description of the system is explained in three parts based on the functionality of each part:

**Exploration.** In the following, the components to support the exploration phase are explained (T1):

- **Number of missing values**: Usually, the first step to understand the missing values is to be aware of the number of missing values. Additionally, it is necessary to know the way the missing values should be handled. In VIVID an interactive categorized stacked bar chart is implemented to show the number of missing values in each wave of the study, each group of features, and each total. As shown in Fig. 3.6(a), bars show the total number of missing values in each wave of the study, and sorted stacks based on the feature's grouping, which are discriminated by the colors, show the number of missing values in

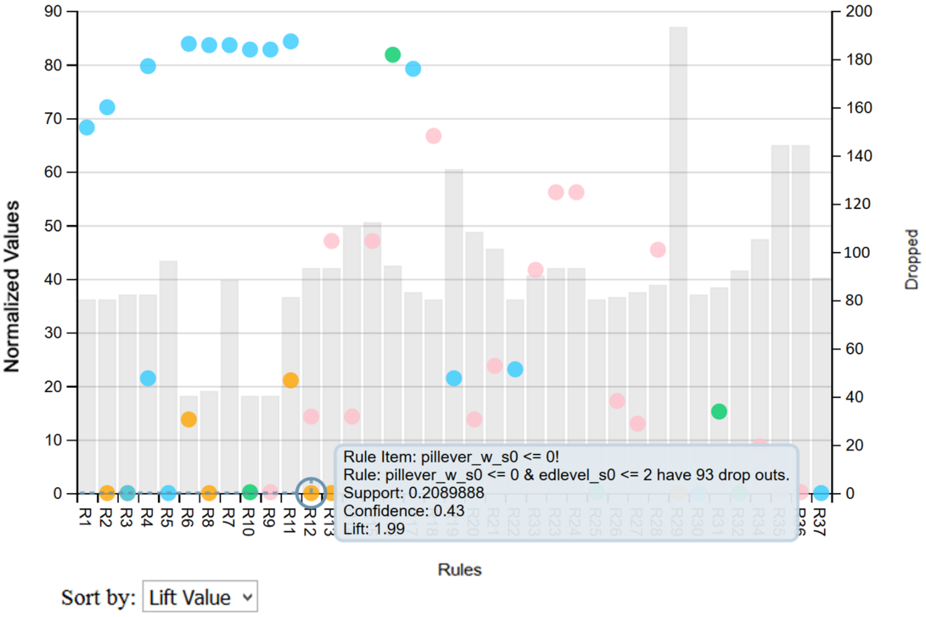


(a) Stacked barchart (b) Missingness map

Figure 3.6: VIVID' user interface consists of: (1) Color-coded categories of features. (2) different charts for exploration of missing values. (3) access to the waves of study. (4) settings for the imputation process. (5) Plots to check the results of imputation. All the plots are visualized in the middle panel of the interface.

each group. The analyst can select a specific feature and the stacks show the quantities only for the corresponding feature in all waves of the study.

- **Location of missing data:** Where the missingness occurs or which participants showed up or dropped out from the follow-up examinations is important to the analyst to create possible strategies for the future to reduce the number of missing values. Additionally, it is helpful to know if there is simultaneous missingness between features. VIVID supports this information by a categorized missingness map. The missingness map provides the most straightforward representation of missing values by showing missing values as empty holes in a matrix representation of the data [38, 43]. As shown in Fig. 3.6(b), the missingness map represents the missing values for a wave of the study, columns represent participants and rows show features that are colored based on the categorization explained in Chapter 2. The rows (features) are sorted based on the specified taxonomy. The missing values are colored black, which reminds the emptiness. A completely black column says that there is a dropout (unit non-response), while a single black cell represents an item non-response. The user can check which participants have missing/observed values w.r.t. a specific group of features, i.e. laboratory values are not measured for participant $x$ because all cells from the laboratory category are black/missing for participant $x$ in the second wave of the study. In the case that some features have missing values simultaneously, the analyst may hypnotize the dependency between features. The main drawback of prior missingness maps in other packages [38] was that it will be cumbersome for large datasets. Thus, VIVID enables the user to adjust the level of zoom by the zooming system. Additionally, hovering the mouse on cell tooltips gives information on the features and participants.



Rule Item: pillever_w_s0 <= 0!
Rule: pillever_w_s0 <= 0 & edlevel_s0 <= 2 have 93 drop outs.
Support: 0.2089888
Confidence: 0.43
Lift: 1.99

Sort by: Lift Value

Figure 3.7: Visualization of association rules. The bars represent the amount of the target feature, i.e. missingness, in each generated rule. The circles are rule items where the coloring is based on the categorization of features and they are normalized between 0 and 100.

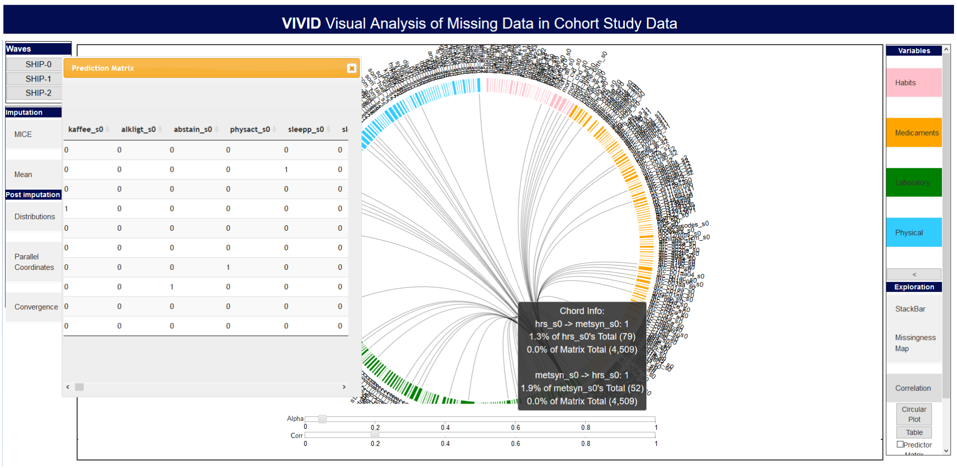
- **Shared characteristics of dropped out participants.** Understanding the common/shared characteristics of participants who drop out of the study may give insight to the analyst to understand the reason or dependencies of missing data. In VIVID, association rules are used to identify these shared features. In the background, classification rules are learned with respect to the participant's status, if the participant dropped out of the study or showed up for follow-up examinations.

To do this, we used the HotSpot package in WEKA [44] and to connect it to R, we employed the RWeka package in R [45]. To learn the classification rules, we generated an additional binary target feature to present the $status$ of participants indicating whether they showed up or dropped out from the second wave of examinations. In the system, the user is able to set the interest of criteria using $status$ as a target feature to generate the rules. For example, if $status = dropout$, it means that the user is interested to find out the subgroups with similar features who dropped out. The minimum support threshold parameter is set to assure the minimum of time the rule is true.

Additionally, the rules' minimum confidence, which represents the minimum length of rules (rule items), can be set by the user. The result of classification rules is shown in a plot to show the rules' features, as presented in Fig. 3.7. Each rule is presented by a bar along with circles, where the bar height shows the number of dropped outs and the circles show rule items (features) - the position of each rule item shows the number of the corresponding features. All features are normalized between 0 and 100. The circles (rule items) are colored based on the conventional colors assigned to the feature categories. It lets the expert identify the interestingness of each rule at a glance w.r.t. the number of dropped outs, contributed features of each category, and length of each rule (subgroup).
By hovering on each rule item, the tooltip gives information on the corresponding rule, i.e. the quality of rule, like lift, support, and confidence values.

**Imputation.** After the exploration phase, VIVID supports the analyst to build the imputation model to preserve the distribution of the data as it should be if the data were not missing (T2). As shown in Fig. 3.6(5), VIVID supports the analyst to set important parameters from the GUI to build the imputation model described in Sec. 3.2.1. These parameters include the maximum number of iterations ($maxit$), the $PredictorMatrix$ and the number of imputed datasets ($m$).

As explained in Sec. 3.2.1, MICE accomplishes the imputation process in an iterative manner. The number of iterations should be enough as the imputations reach a convergence level, i.e. the predicted values of ($m$) different datasets should not fluctuate too much from each other. The user



| kaffee_s0 | alkligt_s0 | abstain_s0 | physact_s0 | sleepp_s0 | sl |
|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 1 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 1 | 0 | 0 |
| 0 | 0 | 1 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 |

Chord Info:
hrs_s0 -> metsyn_s0: 1
1.3% of hrs_s0's Total (79)
0.0% of Matrix Total (4,509)

metsyn_s0 -> hrs_s0: 1
1.9% of metsyn_s0's Total (52)
0.0% of Matrix Total (4,509)

Figure 3.8: The circular graph shows a binary correlation matrix. The features are colored and sorted based on the feature's categorization. A slider enables adjusting the threshold values for adjusting the prediction matrix. When the correlations exceed the thresholds, then arcs show strong correlations. Also, a binary table shows the prediction matrix.

is able to set the parameters $m$ and $maxit$ manually via a text input item. As explained in Sec. 3.2.1, the result of multiple imputation is $m$ imputed datasets. The values for observed data are fixed in all datasets, while they vary for missing data. It is clear that the prediction of the exact values of missing data is infeasible. Thus, the differences between these predicted values express the amount of uncertainty of predicted values.

There are different suggestions to set $m$. The number of $m$ is set depending on the dataset size and the number of missing values [46, 47]. By default in the MICE package, $m$ is set to 5, which in other studies showed that it is enough to give a certain amount of accuracy for the imputations [**?** ]. Thus, in VIVID the user interface is also considered as the default value. Another important parameter is the set of features that is useful to predict the missing values of a feature. These features are passed to the MICE package as an $n * n$ binary matrix, which is called the $predictorMatrix$, where $n$ is the number of features. In the $predictorematrix$ the rows are considered as target (features to be predicted) and the columns as predictors of the target feature, where 1 reflects the corresponding feature and is used in the imputation process of the target feature. In contrast, 0 means that the corresponding feature is not involved in the imputation process of the target feature. In the MICE package, all the features are used to predict a target feature (feature with missing values) by default, which means that a target feature can be predicted by regression over all other features. On the one hand, due to the computation complexities, using the default $predictormatrix$ does not seem efficient when the dataset size is large. On the other hand, the wrong adjustment of $predictormatrix$ may reduce the accuracy of imputation. Thus, VIVID gives suggestions to the analyst to form $predictormatrix$. Besides making suggestions by VIVID, the user should set the predictors carefully to impose the least risk of bias in predictions [48].

To make a trade-off between computations and accuracy of imputation, VIVID calculates an $n*n$ correlation matrix of features to find the relevant features for prediction. The correlation matrix is binarized using a $threshold$ value. In the binary correlation matrix, if the correlation between $feature_a$ and $feature_b$ exceeds the threshold, 1 is written in the position of $a*b$ and $b*a$ of the binary correlation matrix. The correlation between nominal features is calculated by the Pearson correlation coefficient, and between categorical features, the Chi-squared test is used. The significant correlation between features is defined by the p-value of the test. The analyst should find an equitable value as $threshold$ to guarantee the accuracy of imputation as well as the computation time.

To support the analyst to choose the $threshold$, VIVID shows the binarized correlation matrix in a circular graph like a chord chart in which the arcs between elements (features) represent the relationship between features [49]. As shown in Fig. 3.8, the graph's components show features that are colored based on the categorization of the data. A slider located on the button side on a circular graph is used to set the $threshold$ to customize the binarized correlation matrix. By moving the slider, the analyst can check the correlation between features. Having a large number of features may lead to cluttered results in the visualization of information (because of having a lot of arc connections). Thus, by hovering on an item (feature) the connection between other features will disappear. Additionally, by clicking on the arc between $feature_a$ and $feature_b$, tooltips give information on the correlation significance between these two features. Moreover, the data table of the binary correlation matrix is available. This customized matrix can be passed as the predictor matrix to the imputation process.

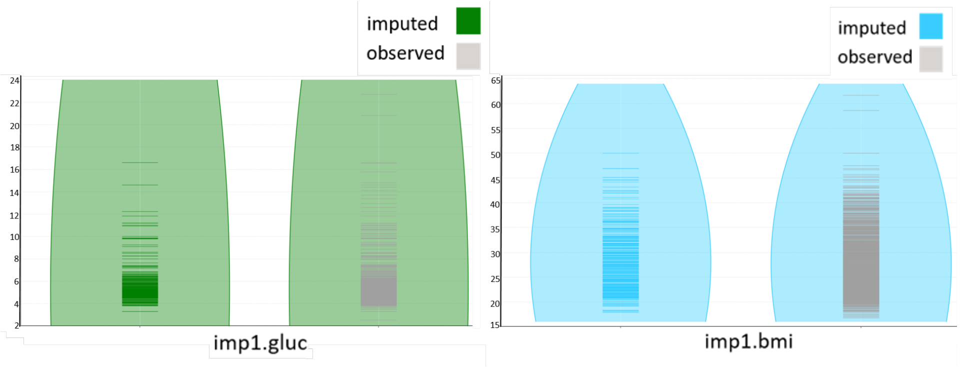


Figure 3.9: Bean plots are used to compare the distributions of imputed and observed values in dense and sparse regions where the lines inside the bean plot represent the values. The colored lines show imputed values and the gray ones display observed values. Beans are colored based on the feature categorization. The left-side plot shows the first imputed dataset for the glucose feature, while the right side shows the BMI feature.

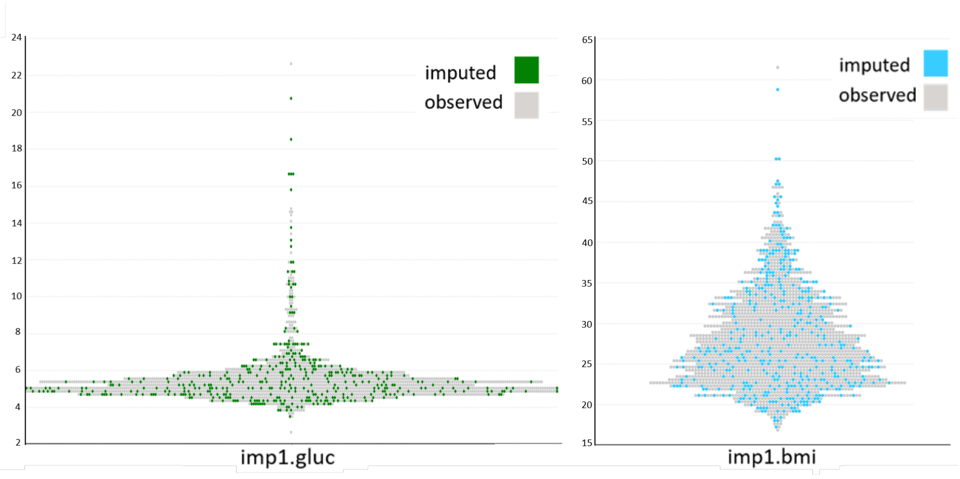


Figure 3.10: The swarm plot displays the distribution of imputed values over observed values in a combined way. The gray points represent observed values and the colored ones (based on the corresponding feature's category) represent predictions. The left-side plot shows the first imputed dataset for the glucose feature, while the right side shows the BMI feature.

**Plausibility of imputations.** After the imputation process, the validity of the results of each $m$ imputed dataset should be checked (T3). The plausibility of the results is usually done by comparing the predicted values over imputed values [50]. Here we considered that the type is MAR, which means, the distribution of observed and predicted values is different. Thus, in the result, it is expected that the minimum and maximum of predicted and observed values are close to each other, while the inner distribution may vary.

VIVID uses bean plots [51] to provide this information. As shown in Fig 3.9, the beans are colored based on the categorization of the features, and inner lines stand for individuals' values, whereas the observed values are grey. It reflects the density and shape of the distribution of the predicted values over observed values side by side for a selected feature.

Additionally, to compare observed and predicted values in a compact way, a swarm plot is implemented. As shown in Fig. 3.10, each circle element stands for a value, whereas the color depicts if it is a predicted value or an observed value.

Last but not least, the difference between predicted values should do not fluctuate too much and the valid results should give a reasonable amount of uncertainty between values in $m$ imputed datasets. To support the validity of this information, VIVID employs parallel coordinates. As presented in Fig. 3.14, $n_{th}$ axis represents the predicted values of $n_{th}$ imputed dataset among $m$ datasets. The lines are predicted values that are colored based on the categorization of the data. The user is able to select an interval from $n_{th}$ to highlight the corresponding predictions in other imputed datasets.

### 3.2.4 VIVID Workflow

In order to describe the functionality of VIVID, the workflow is explained by analyzing SHiP data. The data of female participants of SHiP-0, SHiP-1, and SHiP-2 are loaded into VIVID's components. The reason why we skipped loading men's data is that men's data consists of features that should be systematically missing, e.g. menopause. Thus, it is always better to make a separation between men and women and treat the data separately.

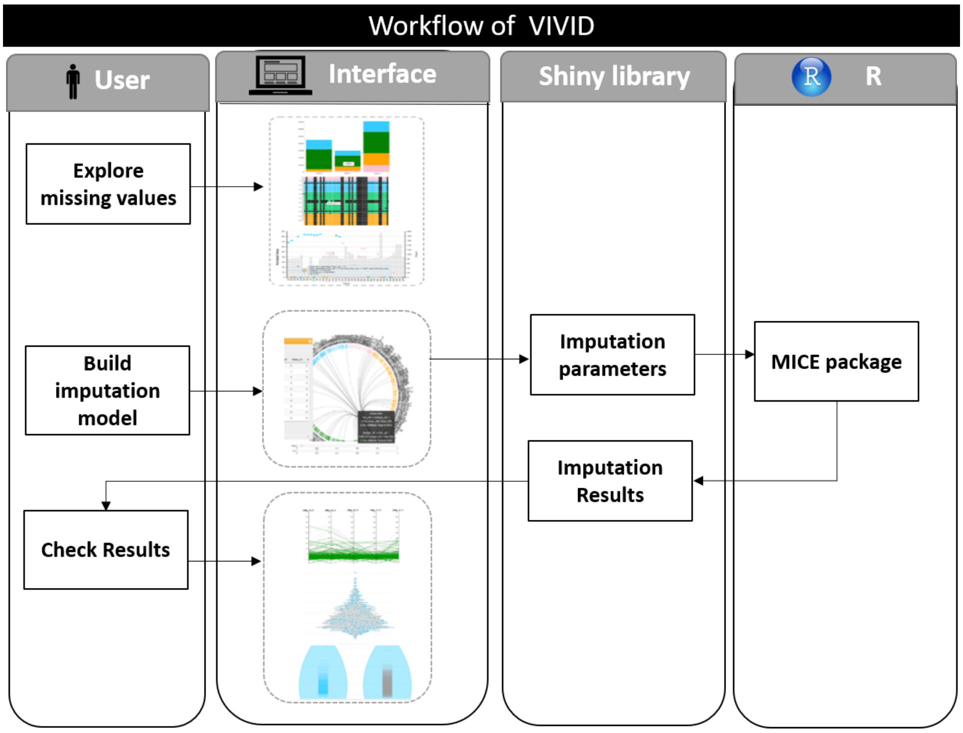


Figure 3.11: VIVID workflow: 1. Exploring the missing values and 2. setting the parameters for multiple imputation. The MICE library in R receives the parameters and 3. runs the imputation process. 4. After imputation, the results will be sent to the analyst to check the quality of imputations via Shiny.

The system is evaluated together with Till Ittermann, who is an experienced statistician working with the SHiP data. The components were explained to him and later he gave feedback on the usability of the system and also suggestions for further improvements.

As explained earlier and shown in Fig. 3.6, the web user interface of VIVID consists of components to navigate the missing values. Thus, in the first step, the expert targets an understanding of missing values by checking the number of missing values in each wave of the study using interactive stacked bar charts. The results show that the last wave of the study (SHiP-2) has the greatest number of missing values and that in all waves of the study features related to laboratory tests are not filled. Moreover, the analyst checked where the missing values occur and where they are unit non-response (drop out) or item non-response cases by missingness maps.

In the next step, the expert generated the association rules to understand the shared characteristics of dropped out of participants (see Fig. 3.7). The results of 37 generated rules from female participants of SHiP-0 showed that women who had never taken birth control tablets and are not well educated dropped out of the study. Additionally, another rule with the highest number of dropouts shows that 193 participants who had never taken birth control tablets dropped out of the study.

In order to start the imputation process, the analyst checks the feasibility of one of the most important parameters, i.e. predictor matrix. As explained earlier, the circular graph shows the binarized correlation matrix between features. The threshold values for the Pearson correlation coefficient and the Chi-squared test were set to 0.2 and 0.05, respectively. These values seem to be sufficient to make the trade-off between computation and accuracy of imputations by giving enough connection between the pairs.

The other parameters $m$ and $maxit$ were left as default to start the imputations. The parameters were sent to the MICE package to start the imputation by Shiny. After completion of imputations, the results were sent back to the user interface for checking the plausibility of the results. In this step, the analyst selected the first imputed dataset and checked the distribution of observed values and imputed values over each other by seeing swarm and bean plots. As shown in Figs. 3.9 and 3.10 for example, BMI and glucose features, the minimum and maximum values were approximately the same as observed values, while the inner distribution slightly varies for both features. The gaps between the observed values were filled for both selected features by the predicted values.

As the last step, the expert should check the amount of uncertainty by comparing the predicted values with each other. To do so, the parallel coordinates are selected for the sample features. The results show that the predicted values in each selected interval did not fluctuate and twisted to each other after 40 iterations.

## 3.3 Evaluation

This section consists of two parts: in the first part, the quality of imputations using multiple imputation is assessed. The second part shows how VIVID makes the trade-off between accuracy and time of computations.

### 3.3.1 Quality of Multiple Imputation

During the imputation process, some questions may arise: what percentage of data needs to be available for the imputation? Does the correlation between features change after the imputation process?

Thus, in the evaluation part, we investigated three aspects of the prediction of missing data in longitudinal epidemiological data using multiple imputations. First, it is analyzed how the accuracy of imputations is influenced by the number of missing values, i.e. with how many missing values the imputation process will lose its accuracy.

Then, it is assessed how the correlation between features may change after the imputation process considering the amount of missing data.

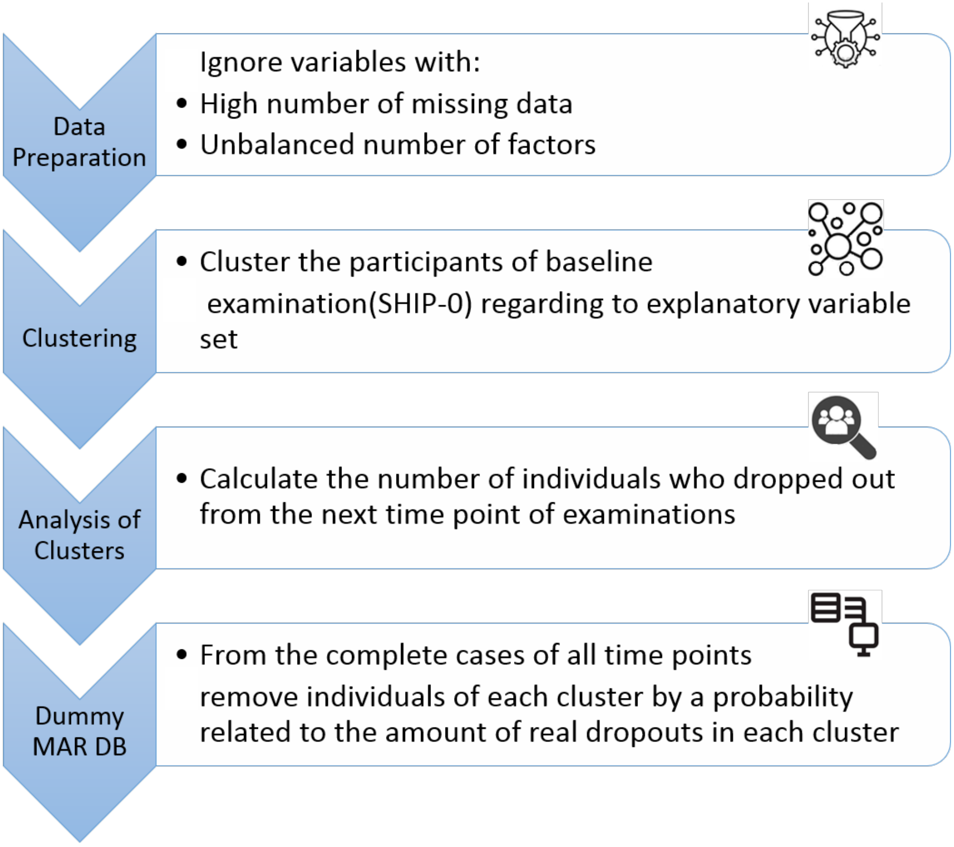
Thus, to evaluate the quality of imputations and check how far the results are from the real values, we created a test MAR dataset (there is a dependency for missing values that are inside the data) using SHiP data. As in some previous studies, it is shown that there is a dependency between dropouts and a set of features ($\rho$) consisting of education level, age, smoking, alcohol consumption-related features, and other disease-related features [52]. This feature set is used to produce the test dataset. The following section explains the steps to build the test dataset (see Fig. 3.12):

1. For preprocessing, exclude features with a high number of missing data and features.
2. Cluster the first wave of examinations (SHiP-0) based on the ($\rho$) feature set. In this way, similar participants with similar dropout dependencies will be grouped together. For this step, the simple k-means clustering was used, which is sufficient to find out balanced clusters, i.e. groups of the same size.

3. Examine the number of dropouts in the next wave of the study (SHiP-1) for each cluster.

4. Combine the two waves of study and create the test dataset by considering only complete cases. Then, give each complete case a probability (considering to which cluster it belongs) to drop out.

5. Remove participants based on the assigned probability to drop out.



Ignore variables with:

- High number of missing data
- Unbalanced number of factors

- Cluster the participants of baseline examination(SHIP-0) regarding to explanatory variable set

- Calculate the number of individuals who dropped out from the next time point of examinations

- From the complete cases of all time points remove individuals of each cluster by a probability related to the amount of real dropouts in each cluster

Figure 3.12: Steps of creating MAR missing data for evaluation.

The clustering method was simple k-means and the number of clusters for creating the test dataset was set to 10. The imputations were compared when 20 vs. 40 percent of participants dropped out. The features glucose and BMI are used here as sample features to show the accuracy of imputations.

**BMI_S1**

| | imp#1 | imp#2 | imp#3 | imp#4 | imp#5 | Real values |
|---|---|---|---|---|---|---|
| min | 16.75 | 16.75 | 16.75 | 16.75 | 16.75 | 16.75524 |
| max | 49.98 | 49.98 | 49.98 | 49.98 | 49.98 | 49.98174 |
| mean | 27.29 | 27.37 | 27.32 | 27.30 | 27.37 | 27.29128 |
| std | 5.056 | 5.184 | 5.093 | 5.143 | 5.168 | 5.105532 |
| var | 25.57 | 26.88 | 25.94 | 26.445 | 26.71 | 26.06645 |
| MAE | 2.330385 | 2.548912 | 2.202117 | 2.375597 | 2.80137 | - |

Table 3.1: Predictions of SHIP1, BMI feature with 20 percent dropouts.

BMI_S1

| | imp#1 | imp#2 | imp#3 | imp#4 | imp#5 | Real values |
|---|---|---|---|---|---|---|
| min | 17.116 | 17.116 | 17.116 | 17.116 | 17.116 | 16.75524 |
| max | 45.34 | 45.34 | 45.34 | 45.34 | 45.34 | 49.98174 |
| mean | 27.48 | 27.32 | 27.22 | 27.38 | 27.35 | 27.29128 |
| std | 5.32 | 5.14 | 5.13 | 5.27 | 5.24 | 5.105532 |
| var | 28.33 | 26.45 | 26.40 | 27.87 | 27.55 | 26.06645 |
| MAE | 2.623 | 2.358 | 2.518 | 2.549 | 2.748 | - |

Table 3.2: Predictions of SHIP1, BMI feature with 40 percent dropouts.

**glucose_S1**

| | imp#1 | imp#2 | imp#3 | imp#4 | imp#5 | Real values |
|---|---|---|---|---|---|---|
| min | 2.54 | 2.54 | 2.54 | 2.54 | 2.54 | 2.54 |
| max | 20.8 | 20.8 | 20.8 | 20.8 | 20.8 | 20.8 |
| mean | 5.308454 | 5.323343 | 5.326667 | 5.325159 | 5.322995 | 5.267594 |
| std | 1.438691 | 1.437309 | 1.446099 | 1.468551 | 1.441486 | 1.310935 |
| var | 2.069831 | 2.065858 | 2.091203 | 2.156641 | 2.077883 | 1.718552 |
| MAE | 0.9678325 | 0.9594089 | 0.968867 | 0.9065025 | 1.029261 | - |

Table 3.3: Predictions of SHIP1, glucose feature with 20 percent dropouts.

In the imputation process, the number of imputations ($m$) is set to 5 and the maximum of iterations ($maxit$) is set to 40.

Tables 3.6 and 3.8 show detailed information of the accuracy of imputations. It shows that for the 20% of dropouts the minimum, maximum, standard deviation and variance are really close to the real values for both BMI and glucose features. Additionally, in predictions of the BMI feature with 40% of missing values, all measurements are close to the real values.

| glucose_S1 | | | | | | |
|---|---|---|---|---|---|---|
| | imp#1 | imp#2 | imp#3 | imp#4 | imp#5 | Real values |
| min | 2.54 | 2.54 | 2.54 | 2.54 | 2.54 | 2.54 |
| max | 16.54 | 16.54 | 16.54 | 16.54 | 16.54 | 20.8 |
| mean | 6.119024 | 5.956976 | 5.767217 | 5.838879 | 6.503082 | 5.267594 |
| std | 2.286906 | 2.178158 | 1.971605 | 2.075997 | 2.614224 | 1.310935 |
| var | 5.22994 | 4.744374 | 3.887228 | 4.309763 | 6.834168 | 1.718552 |
| MAE | 2.889952 | 2.540594 | 2.21114 | 2.40228 | 3.604252 | - |

Table 3.4: Predictions of SHIP1, glucose feature with 40 percent dropouts.

However, the predictions of glucose are less accurate. The reason is probably that in the simulated MAR database we did not have enough participants with a high amount of glucose. The minimum absolute error (MAE) (Eq. 3.2) does not differ too much for the amount of missingness for the BMI feature, while it is getting considerably worse for the glucose feature with 40% of missing data.

The reason is that in reality, the BMI value has a strong correlation with more features like age, sex, weight, and height that are used in the imputation process.

$$MAE = \frac{\sum_{i=1}^{n} |\mathrm{RealValue_i} - \mathrm{Predicted_i}|}{\mathrm{n}} \tag{3.1}$$

Additionally, the following shows how the imputation affects the correlation between features. Figure 3.13(a)(b) shows the correlation between the numeric features age and BMI from the physical status group and serum creatinine from laboratory test groups before and after the imputation for the first imputed dataset. Figure 3.13(c) shows the correlations of real values. From the selected features we can conclude that after the imputation the correlations slightly differ when the amount of missing values is about 20%İn contrast to this, it significantly differs for most pairs of features after the imputation with 40% of dropouts.

In conclusion, multiple imputation works well when there is enough information to complete the predictions based on this information, as with 20% of dropouts the predicted values were close to the real values and the correlation between the features is preserved. In contrast, the accuracy of predictions will considerably reduce having 40% of dropouts.

### 3.3.2 Trade-off between Accuracy and Time

In this section, it is explained if using VIVID helps to have less computation time while having a good accuracy of predictions.

As explained earlier, by default, the MICE package uses all other features in the dataset to predict the missing values of a feature. In the case of having a big dataset, it is time-consuming and sometimes infeasible because of hardware limitations. By sending the predictor matrix suggested by VIVID, we can use only relevant features for the prediction process. To create the predictor matrix, the thresholds of the correlation coefficient are set to 0.2 and the Chi-squared test is set to 0.05. These values give enough relevant predictors for each feature. In the case of giving high threshold values, we may involve more irrelevant features in the imputation process which increases the computation time. In contrast, very low values may skip the relevant predictor feature set, which may lead to inaccurate predictions. By this, the computation time decreased to 14 minutes from 1 hour and 30 minutes. From the accuracy point, Figures 3.14 and 3.15 compares the prediction of the BMI feature in the SHiP-1 dataset using VIVID's suggested predictor matrix and the default predictor matrix.

To make the comparisons more comprehensible, the participants of each group interval are highlighted to see how the predictions of each group are close to the real values. The results show that for the BMI feature the predictions for all groups including underweight, normal, overweight are the same for the default and VIVID's suggested predictor matrix. Figures 3.16 and 3.17 show the predictions for glucose from the laboratory category. For both imputation models, the predictions for normal and pre-diabetes groups are close to each other. The predictions for the diabetic group are less accurate, as they are diverse and far from real values. An explanation for this might be that the number of observed participants for these groups is considerably less than the other groups. In conclusion, the accuracy of the prediction did not reduce by adjusting the predictor matrix, i.e. considering only relevant features to predict a feature, while the computation time reduced considerably.

$$MAE = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} |\mathrm{RealValue_i} - \mathrm{Predicted_i}|}{\mathrm{n}} \tag{3.2}$$

| | |
|---|---|
| Number of participants | 1035 |
| Number of SHiP-0 features | 216 |
| Number of SHiP-1 features | 43 |
| Number of dropouts in SHiP-1 | 204 (≈ 20%) |
| Number of dropouts in SHiP-1 | 421 (≈ 40%) |

Table 3.5: The dummy database with MAR dropout missing values.

**BMI_S1**

| | imp#1 | imp#2 | imp#3 | imp#4 | imp#5 | Real values |
|---|---|---|---|---|---|---|
| min | 18.08 | 19.02 | 16.75 | 18.06 | 16.75 | 16.94 |
| max | 45.34 | 45.34 | 41.98 | 43.35 | 42.94 | 45.104 |
| mean | 26.68 | 27.33 | 26.35 | 26.52 | 27.32 | 26.85 |
| std | 5.40 | 5.32 | 5.29 | 5.11 | 5.17 | 5.37 |
| var | 29.2 | 28.36 | 28 | 26.17 | 26.83 | 28.88 |
| MAE | 3.52 | 3.54 | 3.49 | 2.50 | 3.53 | - |

Table 3.6: Predictions of SHiP-1, BMI feature with proposed predictor matrix by VIVID.

**BMI_S1**

| | imp#1 | imp#2 | imp#3 | imp#4 | imp#5 | Real values |
|---|---|---|---|---|---|---|
| min | 16.75 | 17.31 | 16.75 | 16.75 | 19.61 | 16.94 |
| max | 40.34 | 45.34 | 45.34 | 43.35 | 45.34 | 45.1 |
| mean | 26.74 | 27 | 26.75 | 27.73 | 27.77 | 26.85 |
| std | 4.84 | 4.85 | 5.23 | 4.85 | 4.85 | 5.41 |
| var | 23.44 | 23.61 | 27.39 | 23.55 | 29.34 | 28.88 |
| MAE | 3.48 | 3.45 | 3.52 | 3.56 | 3.50 | - |

Table 3.7: Predictions of SHiP-1, BMI feature with default mode.

### 3.3.3 Summary and Conclusion

In this chapter, we presented a web-based system called VIVID to allow epidemiological analysts to explore, impute, and check the plausibility of the imputations. For the exploration phase, we enhanced conventional visualizations like a missingness map to see where the missingness occurs. To find out the characteristics of dropout participants we employed asso-

**glucose_S1**

| | imp#1 | imp#2 | imp#3 | imp#4 | imp#5 | Real values |
|---|---|---|---|---|---|---|
| min | 2.54 | 2.54 | 3.51 | 3.84 | 3.79 | 3.3 |
| max | 20.8 | 16.54 | 12.94 | 11.12 | 9.3 | 11 |
| mean | 5.46 | 5.46 | 5.22 | 5.03 | 4.92 | 5.11 |
| std | 1.58 | 1.32 | 0.88 | 0.77 | 0.65 | 0.93 |
| var | 2.51 | 1.76 | 0.77 | 0.59 | 0.43 | 1.87 |
| MAE | 0.82 | 0.82 | 0.79 | 0.77 | 0.82 | - |

Table 3.8: Predictions of SHiP-1, glucose feature with proposed predictor matrix by VIVID.

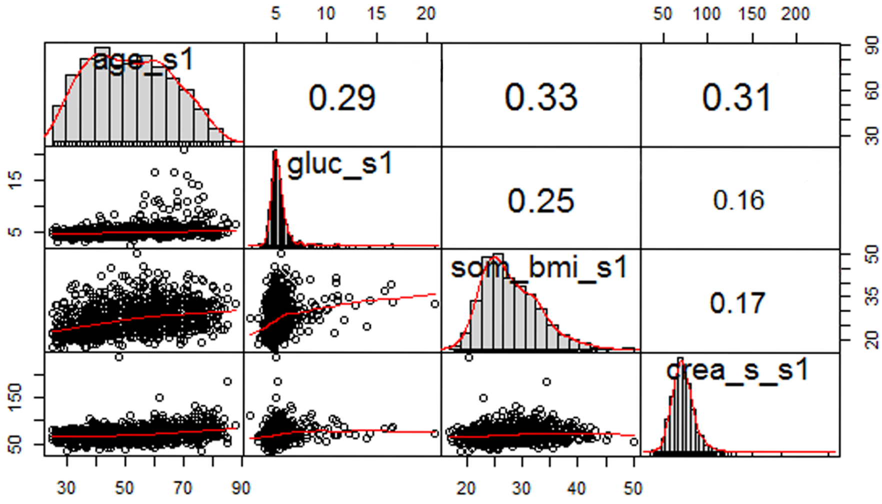
**glucose_S1**

| | imp#1 | imp#2 | imp#3 | imp#4 | imp#5 | Real values |
|---|---|---|---|---|---|---|
| min | 2.54 | 2.54 | 2.54 | 2.54 | 2.54 | 2.54 |
| max | 20.8 | 20.8 | 20.8 | 20.8 | 20.8 | 20.8 |
| mean | 5.45 | 5.36 | 5.45 | 5.32 | 5.23 | 5.27 |
| std | 1.52 | 1.39 | 1.46 | 1.39 | 1.3 | 1.39 |
| var | 3.21 | 1.67 | 2.45 | 1.73 | 0.43 | 1.95 |
| MAE | 0.92 | 0.82 | 0.9 | 0.8 | 0.82 | - |

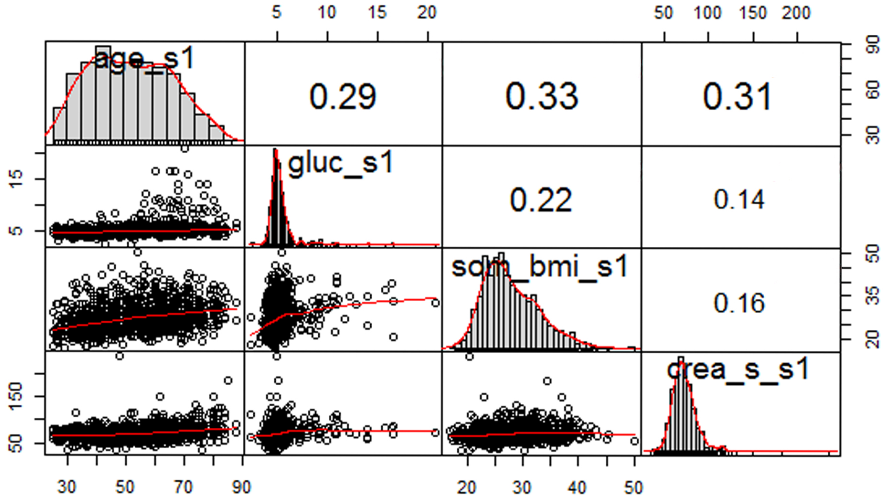
Table 3.9: Predictions of SHiP-1, glucose feature with default mode.

ciation rules. To predict missing values via multiple imputation, VIVID allows the analyst to make the imputation model and set the required parameters and send it to the MICE package in R. Then, it provides visualizations to check the validity of the imputations, e.g. distribution of predicted values over missing values.

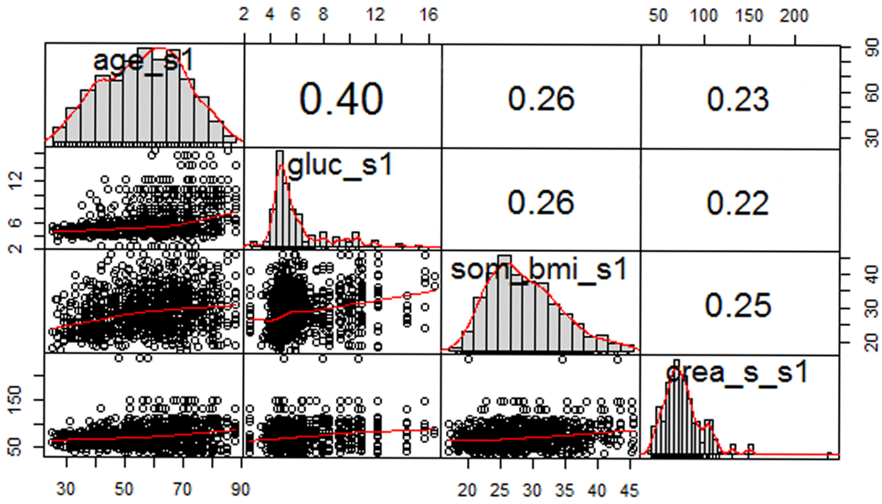
Moreover, an evaluation section is provided to assess the performance of VIVID (to measure how VIVID makes the trade-off between time and accuracy of imputations).



(a) Correlation between real values

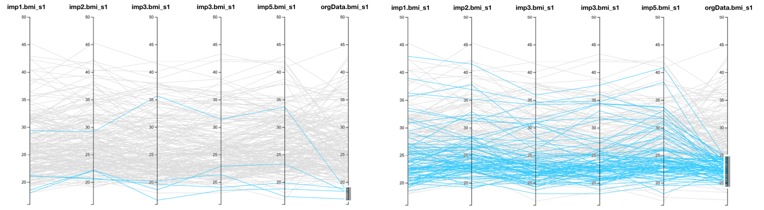


(b) Correlation between variables after imputation with 20% of dropout

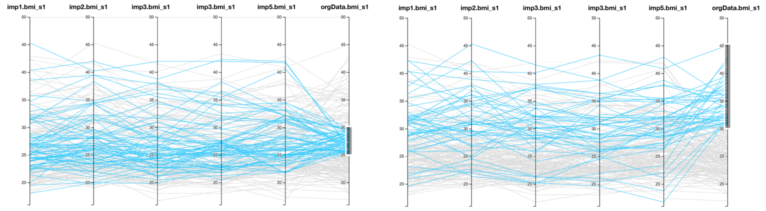


(c) Correlation between variables after imputation with 40% of dropout

Figure 3.13: The correlations between age, glucose, BMI and serum creatinine. (a) and (b) show the correlations for the first imputed datasets with 20 and 40 percent of dropouts, respectively. (c) shows the correlations of real values.

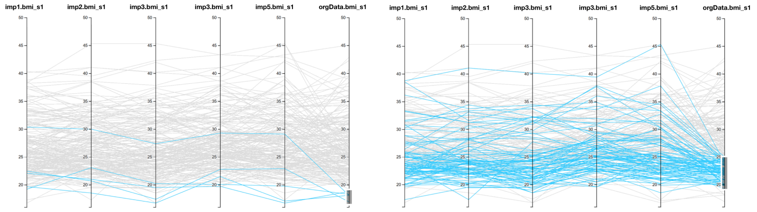


(a) Underweight (b) Normal

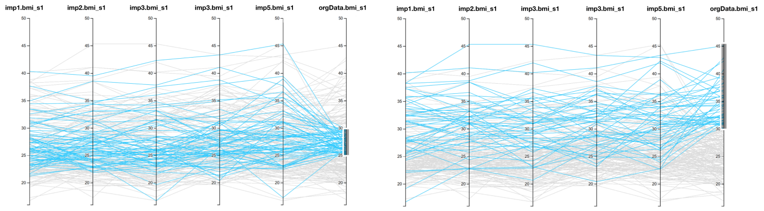


(c) Overweight (d) Obese

Figure 3.14: Predictions of BMI, from the physical status category, with 20 percent dropout. (a) Underweight participants BMI<18.5, (b) Normal weight participants 18.5<BMI<25, (c) Overweight 25<BMI<30, (d) Obese BMI>30.

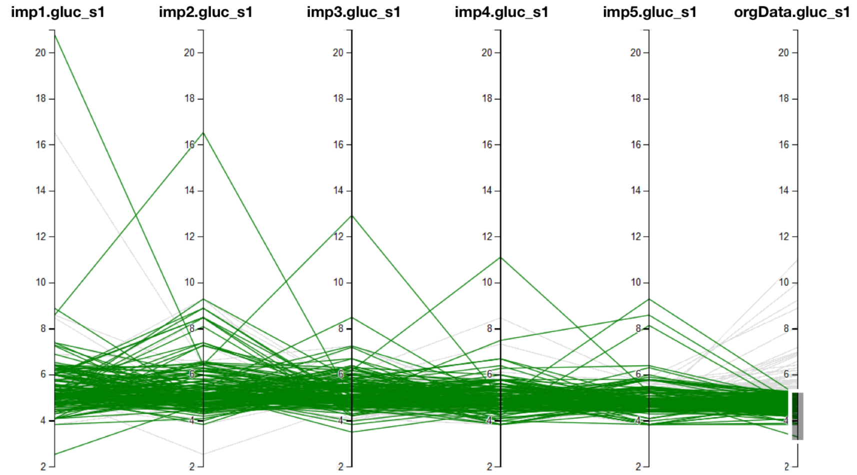


(a) Underweight (b) Normal

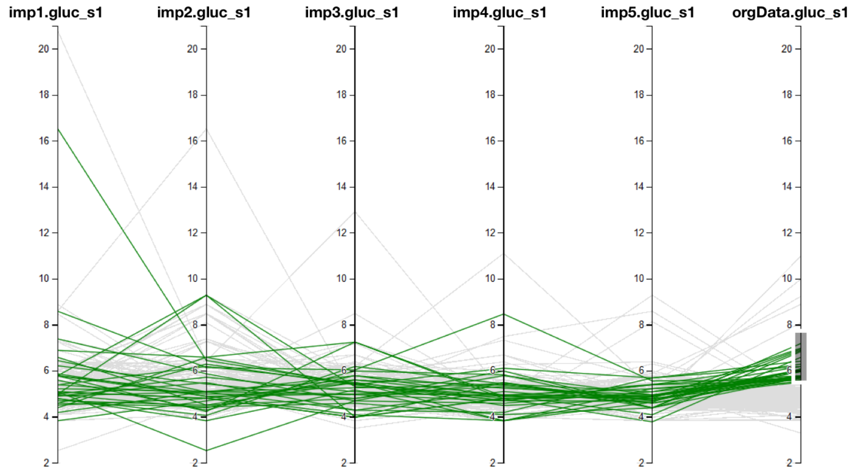


(c) Overweight (d) Obese

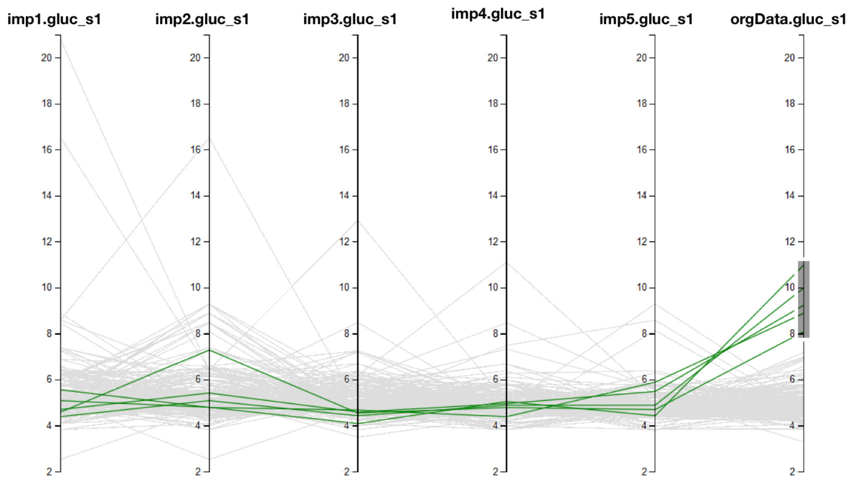
Figure 3.15: Predictions of BMI, from the physical status category, with default prediction matrix. (a) Underweight participants, (b) Normal weight participants, (c) Overweight, (d) Obese.



(a) Normal level of glucose

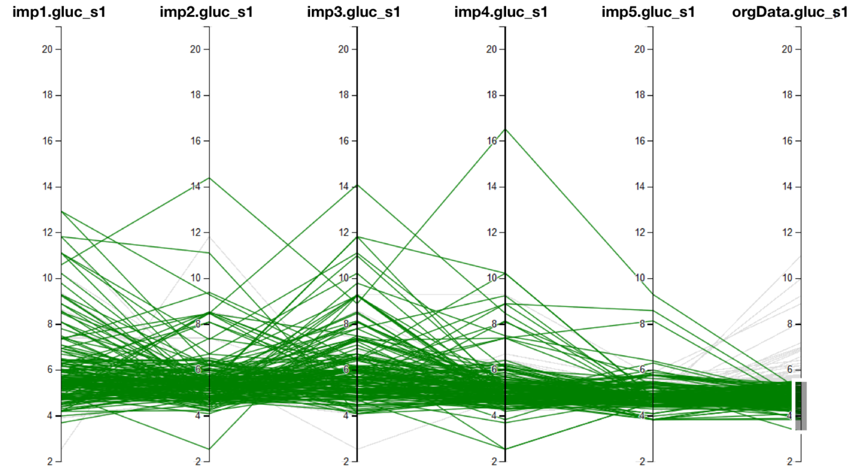


(b) Pre-diabetes level of glucose

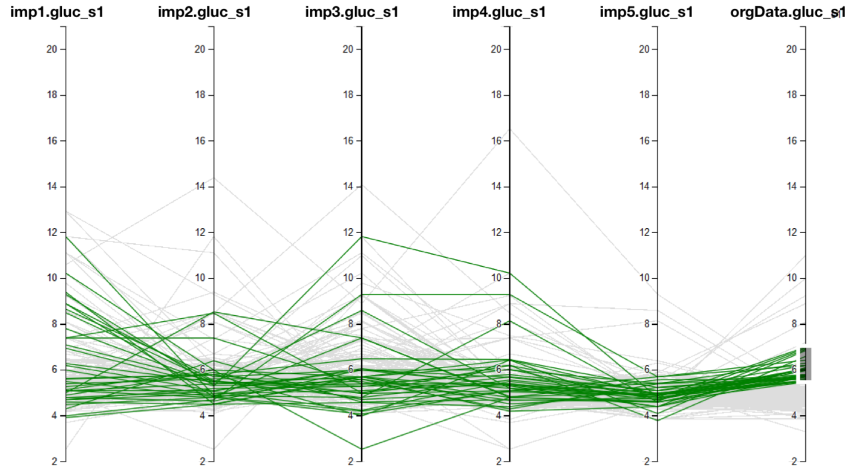


(c) Diabetes level of glucose

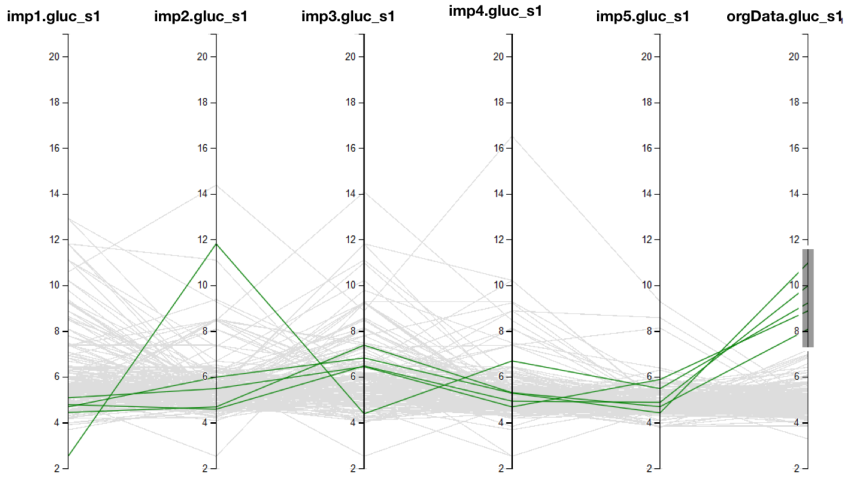
Figure 3.16: Predictions of glucose, from the laboratory category, with 20 percent dropout. (a) Normal up to 5.5, (b) Pre-diabetes between 5.5 to 7, (c) Risk of diabetes above 7.



(a) Normal level of glucose



(b) Pre-diabetes level of glucose

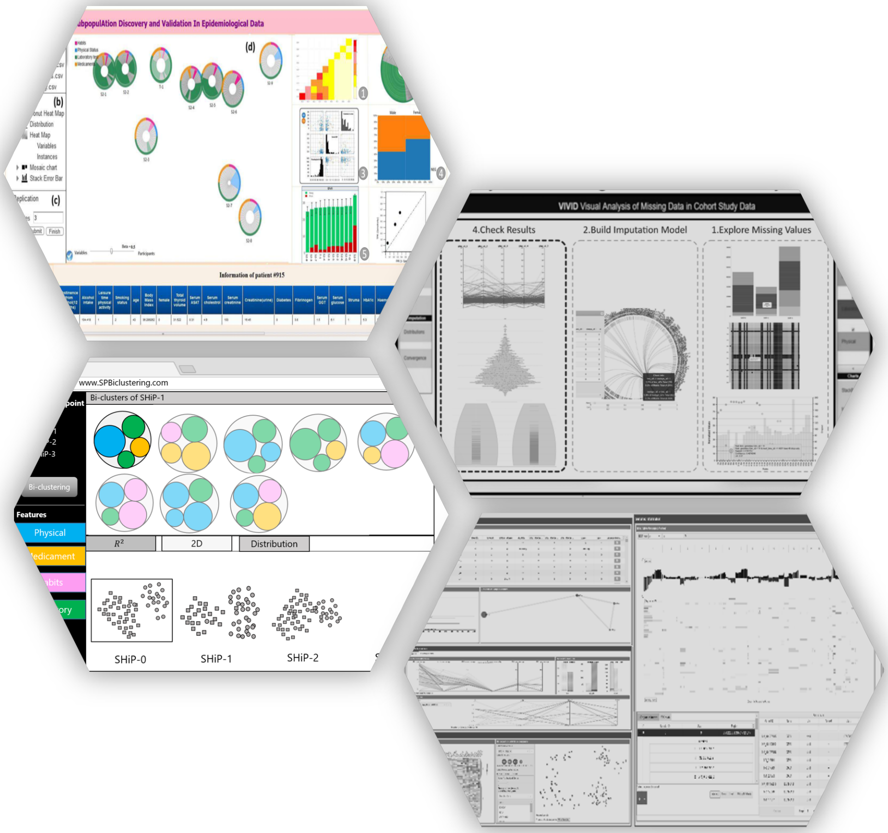


(c) Diabetes level of glucose

Figure 3.17: Predictions of glucose, from the laboratory category, with the default imputation model. (a) Normal up to 5.5, (b) Pre-diabetes between 5.5 to 7, (c) Risk of diabetes above 7.

# 4

# Epidemiological Sub-cohort Discovery



One of the main goals of epidemiology studies is to investigate the causes of diseases by discovering the risk factors which are related to a population's lifestyle, genetic events, living environment, and socio-demographic factors. Thus, there is a need to identify the high-risk and low-risk sub-cohorts which discriminate from the whole cohort. In data mining, clustering is an unsupervised approach is deployed for grouping a cohort. In a cluster, the individuals are related to each other through a biological or social characteristic or they have a relationship with a special event [53].

## 4.1 Clustering

The general description of clustering is described in Chapter 2. In the rest of the document, the subjects and participants have the same meaning. The term clustering refers to the grouping of similar subjects in a dataset. Usually, the input of a clustering algorithm is a dataset where the column forms the features and each row forms a subject. In other words, a subject can be described as a point in the multidimensional space. The subjects are grouped based on a specific distance factor, such as Euclidean w.r.t. their features. Nowadays, the advancements in data collection technologies lead to larger, high-dimensional, and consequently more complex data. With the additional complexity of the data, the feasibility of the traditional clustering algorithms becomes questionable since to measure the similarity of the subjects they consider all dimensions in the distance function while many features may not be representative for the distance of the two specific objects. In high-dimensional complex data, it is more likely that data is noisy. Thus, by involving all features in the clustering the results are possibly incorrect as many clusters maybe are hidden in subspaces. The other issue with traditional clustering of high dimensional data is the curse of dimensionality. This term means that with the increasing number of features in a dataset, the distances between subjects become vague. It means by adding more features, the subjects spread out from each other until the distance between them becomes equal. Figure 2.9 displays the issue of the curse of dimensionality. Thus, the clustering on data with a lower number features of features is more meaningful for the analysis.

Although some techniques like feature selection can handle this problem, they usually only consider a fixed subset of dimensions in the clustering procedure, while different objects maybe group-able in different subspaces. Many times, different objects are related to each other in different subspaces.

**Subspace Clustering**

Subspace clustering is a technique that considers all dimensions in the clustering process and the algorithm finds similar objects that can be grouped together in an overlapped subsets of dimensions. There are a number of works that review the (subspace-)clustering approaches. In fact, subspace clustering is the enhanced version of the feature selection

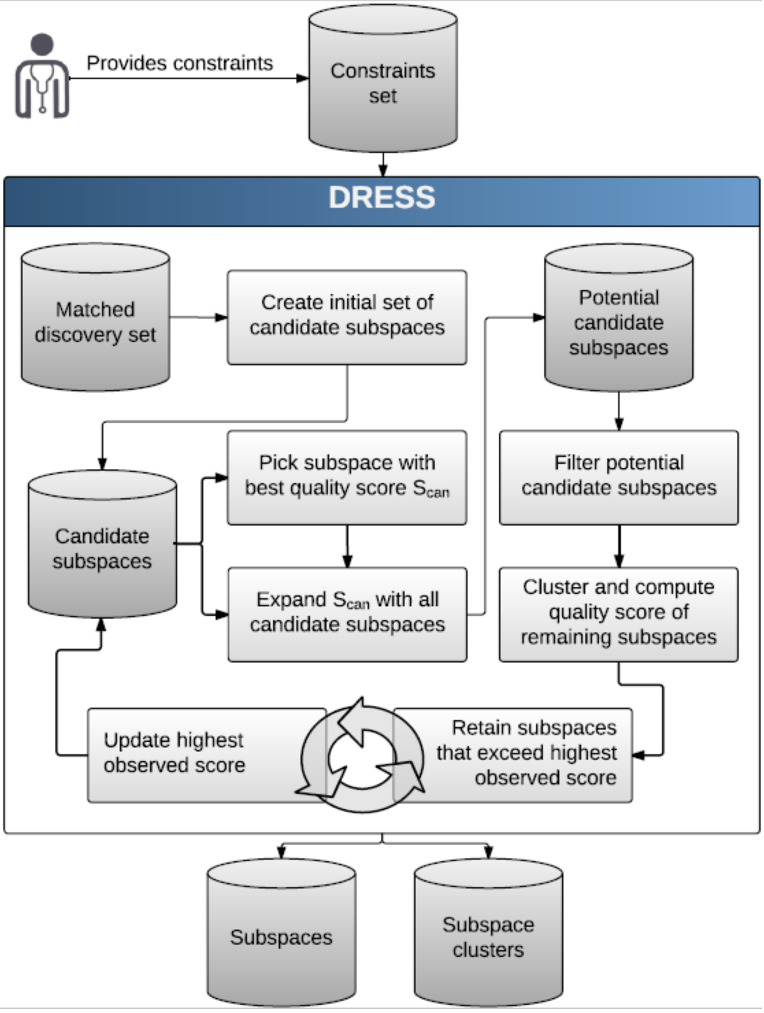

Provides constraints
Constraints set
DRESS
Matched discovery set
Create initial set of candidate subspaces
Potential candidate subspaces
Candidate subspaces
Pick subspace with best quality score $S_{can}$
Filter potential candidate subspaces
Expand $S_{can}$ with all candidate subspaces
Cluster and compute quality score of remaining subspaces
Update highest observed score
Retain subspaces that exceed highest observed score
Subspaces
Subspace clusters

Figure 4.1: The workflow of the DRESS algorithm [54].

and traditional clustering technique. Its main aim is to find groups of objects in a different subsets of dimensions, i.e. subspaces, without eliminating some dimensions or objects. Overall, it has two main steps: first, feature selection and the second step is to attempt the clustering technique. In other words, subspace clustering needs a search method to evaluate each combination of subspaces to apply the clustering.

In this work , we used a constraint-based clustering technique to find sub-cohorts associated with specific outcomes using labeled participants. Although labeling of participants is not always feasible because of expensive medical examinations, semi-supervised constraint-based subspace clustering approaches overcome this limitation. They find clusters in subspaces using a small amount of background knowledge of participants' characteristics which are associated with medical outcomes.

DRESS (*D*iscovery of *R*elevant *E*xample-constrained *S*ub*S*paces) is a constraint-based clustering technique which is particularly helpful to find subspace clusters of participants by incorporating experts knowledge [54].

To determine similar participants w.r.t. the terms of must-link and not-link, the*DRESS* algorithm uses a few numbers of constraints defined by the medical experts. A must-link constraint defines that two subjects in a must-link relation should be located in the same cluster, while two subjects with not-link relation should not be present in the same cluster. For example, to find subspace clusters of participants that are highly correlated with a disease like fatty liver, the expert uses a bit of background knowledge on the disease to define must-link constraints between participants with the same disorder, i.e. positive and negative fatty liver. The not-link constraints between participants can be defined in a similar way. As presented in Fig. 4.1, the workflow of *DRESS* is as follows [55]:

1. *DRESS* starts with a quality scoring of each subspace of cardinality one. Initially, these subspaces constitute the candidate set of subspaces. The subspace quality is scored by considering the distance between must-link and not-link constrained participants in the respective subspace as well as the proportion of satisfied constraints to all constraints. For the respective subspace, a must-link constraint is satisfied if both constrained participants lie within the same cluster

and a not-link constraint is satisfied if both participants are members of different clusters.

2. *DRESS* iteratively picks the best-scored subspace $S_{can}$ and merges it with all remaining subspaces in the candidate set. To reduce complexity, the resulting subspaces are filtered by a reduced quality criterion (faster to compute than the full quality), i.e. if the calculated quality part is lower than for the original subspaces, the merged subspace is not further considered. For all subspaces that satisfy the filter criterion, the full quality is calculated, which involves a density-based clustering with DBSCAN [56] where parameters are automatically determined [57]. As soon as the quality of a subspace exceeds the highest yet observed quality $q_{best}$, *DRESS* retains it as a candidate subspace for further extension, updates $q_{best}$ and stores all contained clusters.

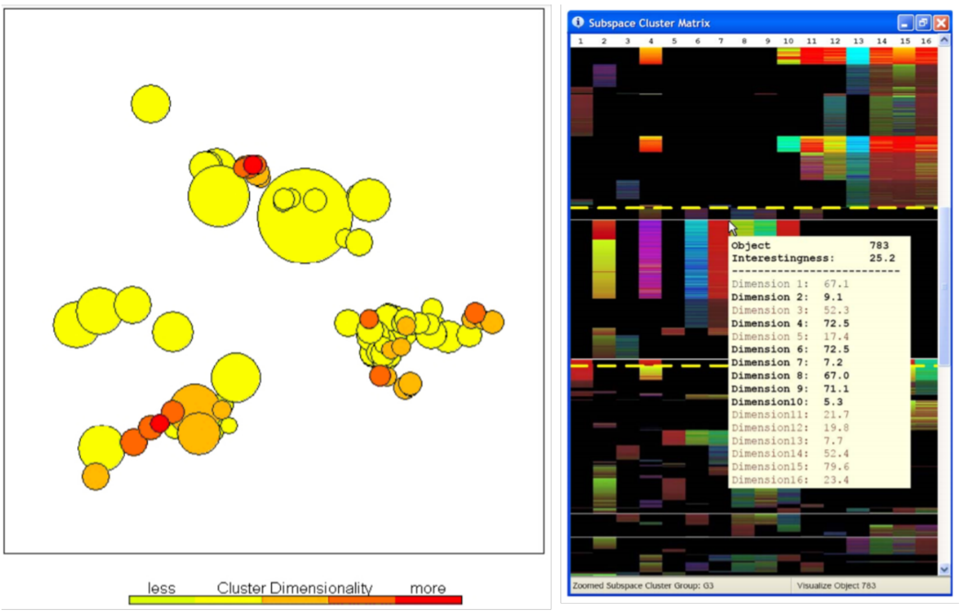
3. At the end of an iteration, $S_{can}$ and all merge candidates that led to a new $q_{best}$ are removed from the candidate set. *DRESS* terminates when the candidate set is empty and returns a ranking of subspaces and their associated clusters.

   Often, subspace clusters that have a high quality according to an *interestingness measure* are very similar, e.g. they differ only in one dimension. To enable the analysis of a representative subset of subspace clusters, it is helpful to analyze such relations and show them graphically, e.g. as a hierarchy visualization [58].

## 4.2 Related Works in VA

In this section, an overview of the analysis and visualization of high-dimensional data, e.g., cohort study data, using subspace clustering is given.

The work of Assent et al. is one of the early works on the visualization of sub-pace clustering results [59]. They proposed a system for the Visual Subspace Clustering Analysis called VISA. As shown in Figure 4.2, the results are shown in two levels, globally and in detail. In the global view (Fig. 4.2(a)) the subspace clusters are visualized as circles, whereby multidimensional scaling (MDS) the distance between circles represents the



```
Object                 783
Interestingness:       25.2
------------------------------
Dimension 1:  67.1
Dimension 2:  9.1
Dimension 3:  52.3
Dimension 4:  72.5
Dimension 5:  17.4
Dimension 6:  72.5
Dimension 7:  7.2
Dimension 8:  67.0
Dimension 9:  71.1
Dimension10:  5.3
Dimension11:  21.7
Dimension12:  19.8
Dimension13:  7.7
Dimension14:  52.4
Dimension15:  79.6
Dimension16:  23.4
```

(a) (b)

Figure 4.2: VISA framework provides visualization of subspace clusters [59]. The left side panel shows an overview of the subspace clustering result w.r.t. overlaps between cluster features and members. The right side panel shows a detailed view of subspace clusters using heatmaps.

similarity between clusters w.r.t. the overlapped features and members of clusters. In the global view, the size of each cluster is encoded as the size of the circle, while the color-coding is representative of the number of involved features for each cluster. Another view shows more detailed information on the quantities of each cluster (Fig. 4.2(b)) by a matrix-like visualization of clusters, where the colors stand for the values of the involved features. Although the visualization gives insight to the user on the results of subspace clustering, it is not a scalable solution because clusters will be overlapped in the global view. Moreover, because the size of clusters is represented as the size of the circle, the big clusters will hamper other adjacent clusters.

In the same year, Achtert et al. introduced a subspace clustering algorithm called DiSH to find the hierarchy of subspace clusters [58]. The algorithm finds clusters with different numbers of features, members, and shapes. The simple tree visualizations show the relationship between the hierar-

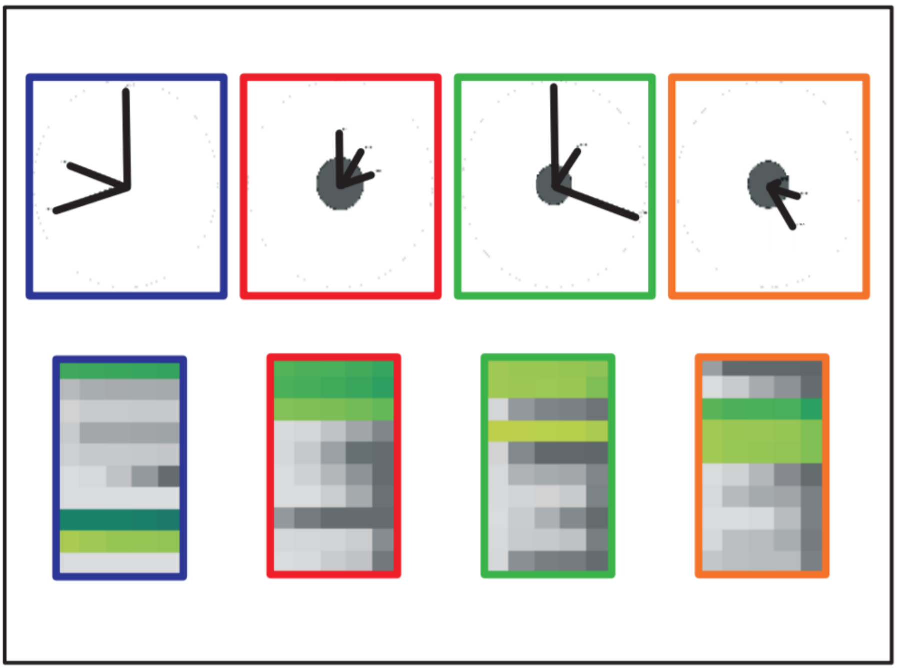


Figure 4.3: ClustNail system gives insight to the analyst on the subspace clustering [60]. The top row shows the clusters where the radius of the circle represents the size of the clusters, while the spikes give information on involved features of the corresponding subspace cluster. The bottom row shows the information on the values of features by heatmaps.

chy of subspace clusters. Later, Tatu et al. presented a system called ClustNail for the visualization of subspace clusters [60]. ClustNail uses a combination of techniques and new visualizations to show the results. The overall similarity of the subspace cluster is performed by an ordering function. Like the previous work, as presented in Fig. 4.3 the clustNail shows clusters by circles along with spikes which are representative of the involved features. A fixed position is given to each feature in each cluster. The length of spikes represents the importance of the corresponding feature regarding its variance; high variance features get a larger length. The heatmaps present the detailed information of each individual cluster.

Hund et al. presented SubVis as a multiple coordinated views system to explore and analyze the results of subspace clustering on large-scale medical data [61]. A global overview panel displays the similarity of clusters using the MDS technique. In addition to the presentation of the features in pre-

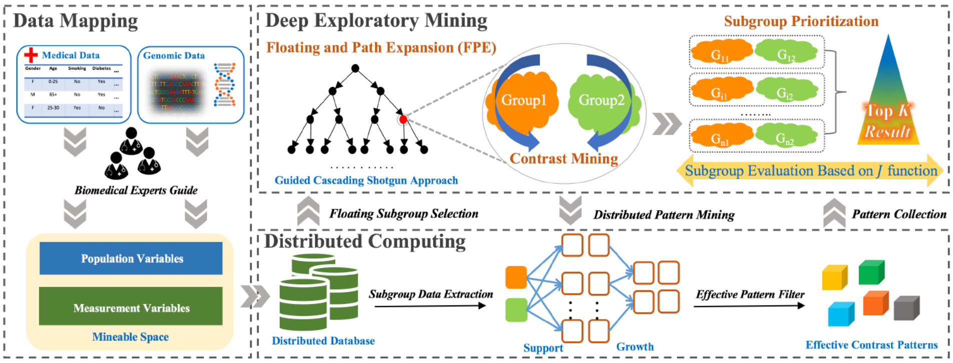


Figure 4.4: The overall workflow of the system proposed by [63] using Exploratory Data Mining for Subgroup Cohort Discoveries and Prioritization. The workflow consists of three parts: data mapping, deep exploratory mining and distributed computing.

vious works, SubVis shows some statistical information of each cluster by conventional visualizations such as scatter plots and bar charts.

Recently, [62] proposed a system to support functionalities for data prepossessing of small electronic health records of patients, i.e. data cleaning and transformation by removing features with distributions which may cause problems for further investigations and also handling missing entries. Then, their system provides analysis on a cleaned dataset for exploration and cohort discovery. It enables the expert to compare sub-cohorts from different statistical levels.

Liu et al. [63] presented an exploratory data mining for sub-cohort discovery. The overall workflow of their technique consists of three steps (see Fig. 4.4). The first step is data mapping, where the expert user maps the raw data to meaningful and formatted data. Secondly, the formatted data is fed to the proposed exploratory mining process and discriminated subgroups and their feature patterns are identified. Thirdly, the discovered sub-cohorts will be evaluated and ranked based on some defined constraints.

## 4.3 Data for Analysis

In this chapter, the Study of Health in Pomerania (SHiP) is used for analysis [64]. The same as other sources of epidemiological data, the information is acquired via interviews of participants and gaining information on sociodemographics, lifestyle, and medicament. Moreover, some information is collected based on physical examinations, for example, BMI and blood pressure status, while some information is based on laboratory examinations, e.g. diabetes status and information on liver functionalities. Additionally, some information is extracted from the medical images. Combing all these features leads to a heterogeneous (categorical and numeric features), high-dimensional dataset [65].

As shown in Fig. 4.5, the SHiP study is carried out in Western Pomerania. The whole study consists of two independent sets of data. The first collection of the first dataset started from 1997 to 2001, i.e. SHiP-0. The second wave of SHiP-1 was from 2002 to 2006, and the third wave was between years 2008 and 2012. The first study was accomplished by SHiP-3 between 2008 and 2012. The main aims of the study were frequent diseases like fatty liver, breast cancer, and back pain. As explained in the previous chapter, the study was initially started with 4308 participants, while within each next wave of study the number of participants reduced dramatically. Thus, a new study started in the same region with new participants, called SHiP-TREND.

In this work, fatty liver disease was targeted. Information on the status of the liver was extracted and annotated by biomedical experts. The participants having more than 10% of saturated fatty on their liver are considered as positive for fatty liver. In this work, SHiP-2 data was used for the analysis, while SHiP-TREND-0 data was used for validation purposes.

In the data sample for the analysis, SHiP-2 participants were significantly older than TREND-0 participants, the proportion of women was higher and the distribution of the outcome (fatty liver status) differs slightly. To receive reliable quality estimates when validating one cohort's findings on another, a nearest neighbor propensity score matching [66] on age, sex, and the outcome was applied. After matching, each 694 of the SHiP-2 and TREND-0 participants remain, with an age of 55.5 ± 12.6 years, 46.5 % men, and 21.5 % fatty liver positive.

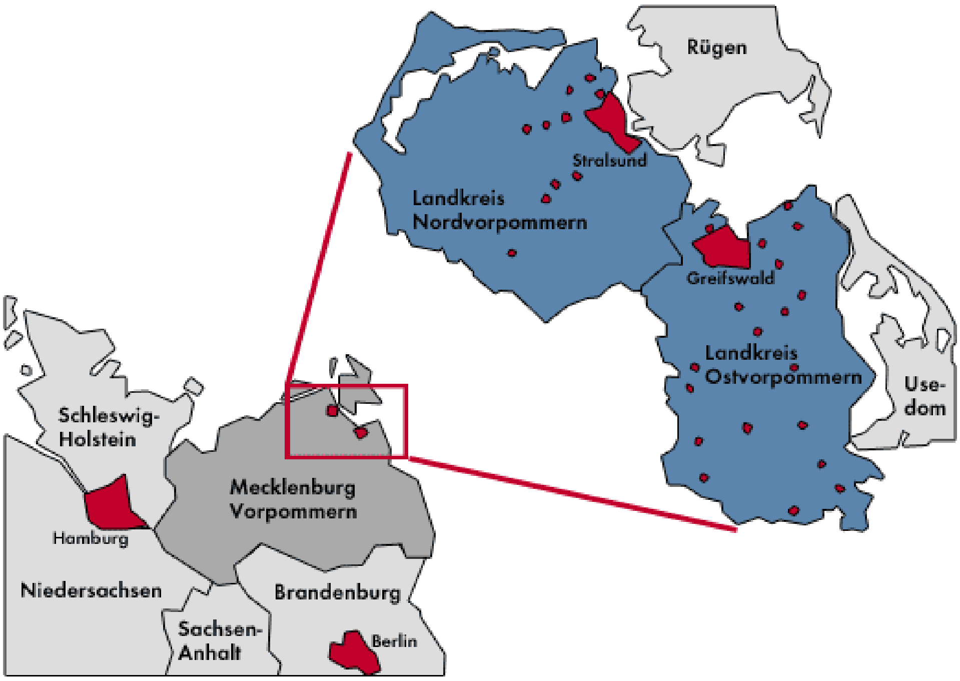


Figure 4.5: Study of Health in Pomerania region (www.wikipedia.org)

## 4.4 Subpopulation Discovery in Cross-Sectional Data

This section is mainly based on

- Alemzadeh, S., Niemann, U., Ittermann, T., Völzke, H., Schneider, D., Spiliopoulou, M., Bühler, K. & Preim, B. (2020, February). Visual analysis of missing values in longitudinal cohort study data. In Computer Graphics Forum (Vol. 39, No. 1, pp. 63-75).

### 4.4.1 Task Analysis

Generally, the tasks that should be covered by S-ADVIsED consist of:

- **Exploration and analysis of discovered subspace clusters** are necessary for the user to get insight on each cluster w.r.t. the relationship of involved features and enable the user to compare different subspace clusters regarding different criteria. It allows the expert to assess the interestingness of each sub-group of patients.

- As subspace clusters have arbitrary shapes, there should be functions to enable the expert to **transform the subspace clusters to sub-cohorts** by describing them in a rectangular shape. The expert should be able to trust the results of subspace clusters by **validation of the findings.**

The sub-tasks are described as follow:

T1 **Involved features:** The expert needs to have an overview of the involved features, i.e. number and type in each subspace cluster. Generally, medical experts are more interested in subspace clusters with a low number of involved features. The knowledge derived from subspace clusters should be transferred to clinical practice and rules, for example, contribute to the prevention, diagnosis, and treatment of diseases. Additionally, subspace clusters with a low number of involved features are less prone to overfitting (significant), as they follow the principle of scientific parsimony.

T2 **Cluster Size:** Having information matters to the expert to know how many participants have the same characteristics. In a medium-sized study like SHiP, each subspace cluster should have at least 5 % of the members of all study participants to support the evidence of statistical significance.

T3 **In-depth Overview:** A clear overview for each subspace cluster should be provided to give insight to the user on the distribution of involved and non-involved features. This allows knowing the cluster compactness. It means that participants who belong to one subspace cluster should be similar to each other with respect to their involved features. For example, when BMI is an involved feature in one subspace cluster, then it is expected that all individuals have close BMI values, i.e. between 20 and 24.

T4 **Feature and individual overlaps:** It is necessary to have a view of the similarity between clusters w.r.t. their shared features and individuals in subspace clusters.

T5 **Deviation of sub-cohorts vs. whole cohort:** Gives information to the experts to evaluate the interestingness of sub-cohorts by comparing them to the whole cohort by knowing how strongly sub-cohorts deviate from the whole population w.rt. a specific target features, e.g. fatty liver. It helps to identify high-risk sub-cohorts.

T6 **Distribution:** The user should be able to assess the relationship of involved features and target feature by an overview of the distribution on the corresponding cluster w.r.t. the involved features.

T7 **Variability of features:** It might be interesting for experts to investigate the reason for the incorporation of high variance features in the results of subspace clustering. Subspace clustering algorithms typically minimize the sparsity of data by ignoring features with higher variance.

T8 **Cluster description:** Many subspace clustering algorithms are applicable to find clusters with any shape, however, to define them as sub-cohorts they need to be specified within intervals in the form of hyper-rectangles. Thus, clusters should transform from arbitrary to the rectangular shape.

T9 **Validation:** The identified sub-cohort as the result of the subspace clustering should be validated to trust the findings.

### 4.4.2 S-ADVIsED System

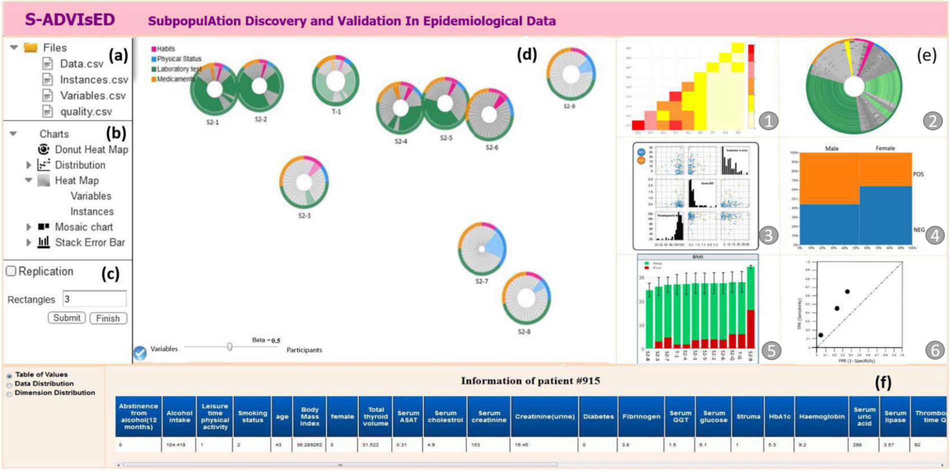
The overall functionalities of S-Advised consists of two parts:

- Exploration of subspace clusters/sub-cohorts
- Cluster description and validation of findings

The following section explains each functions and its components in detail:

#### Exploration of subspace clusters/sub-cohorts

The S-ADVIsED consists of a fixed global view of all sub-space clusters/sub-cohorts and a detail view which shows the desired information for the exploration of:



Information of patient #915

| Abstinence from alcohol(12 months) | Alcohol intake | Leisure time physical activity | Smoking status | age | Body Mass Index | female | Total thyroid volume | Serum ASAT | Serum cholestrol | Serum creatinine | Creatinine(urine) | Diabetes | Fibrinogen | Serum GGT | Serum glucose | Struma | HbA1c | Haemoglobin | Serum uric acid | Serum lipase | Thrombo time Q |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 104.418 | 1 | 2 | 43 | 36.289262 | 0 | 31.522 | 0.31 | 4.9 | 103 | 16.45 | 0 | 3.8 | 1.5 | 6.1 | 1 | 5.3 | 9.2 | 298 | 3.67 | 92 |

Figure 4.6: The S-ADVIsED user interface consists of 5 panels: (a) files for analysis, (b) plot, (c) validation/ replication panel, (d) global overview of subspace clustering result, (e) detail view, (f) statistical info of subspace clusters.

- **Cluster overview:** This fixed view gives an overview of all subspace clusters and their characteristics. Each cluster is represented as a donut chart which encoded clusters' properties (Fig. 4.7). Moreover, this view represents the distance of clusters (T2, T1 and T4).

  **Similarity:** The similarity of clusters is represented in the global view in the terms of the overlap of members or features. In this way, clusters that have more participant overlaps are closer to each other. To do this, first, the distance between clusters should be calculated, then according to the similarity matrix, they should be projected in 2D space (T4).

  For the first task, to measure the similarity between subspace cluster $SC_i$ and $SC_j$ in subspaces $S_i$ and $S_j$, the same equation as in the work of [59] is used.

  According to Equation 4.1, the fraction of overlapped features is reflected in the first part of the equation, while the second part of the equation represents the fraction of overlapped participants between $SC_i$ and $SC_j$. The importance of the participant overlaps or feature overlaps are assigned by giving a weight to the equation by the $\beta$ value. The value of $\beta$ is between 0 and 1, whereas 0.5 depicts participant overlaps, and feature overlaps have no priority on each other. In the same way, if in the projection of clusters the overlap between features has a priority to the overlap between participants, a higher value of more than 0.5 should be assigned to the $\beta$. In the global view, the user can set the value of $beta$ by a slider.

$$dist_{SC_i,SC_j} = \beta\left(1 - \frac{|S_i \cap S_j|}{|S_i \cup S_j|}\right) + (1-\beta)\left(1 - \frac{|C_i \cap C_j|}{min\{|C_i|, |C_j|\}}\right) \quad (4.1)$$

After calculation of the similarity matrix, the clusters are projected in 2D space by multi dimensional scaling (MDS) dimension reduction [67].

**Sectors:** The size of the sectors is equal to the size of the feature. Thus, each sector represents a feature, as the size of the sector depicts the variance of the corresponding feature. Thus, a feature with higher variability has a bigger sector (T7).

**Size:** The size of a cluster is represented as the size of the corresponding donut chart. Hence, the cluster with the highest number of individuals has the biggest radius (T2).

**Color coding:** The colors of sectors represent the categorization of features. The corresponding sector of an involved feature is colored by the assigned colors to each category discussed in Chapter 2, while non-involved features are gray. Thus, the user can assess at a glance which group of features participated in each cluster (T1).

**Interactions:** The user is allowed to see the status of a specific feature by the linking and brushing technique. A selected feature in a cluster will highlight the corresponding feature in all other clusters in the global view.

- **Plots** As presented in Figure 4.6 (e), this view shows the desired information for the analysis of the subspace clustering result. The plots are accessible via a plot panel located on the left side, see Fig. 4.6 (b). This plot consists of:

  **Donut heatmap:** The donut heatmap provides more detailed information on each selected cluster. It has the same characteristics as the donut plot in the global view, whereas the heatmap shows the information on the values of each feature. As shown in Fig. 4.8, the

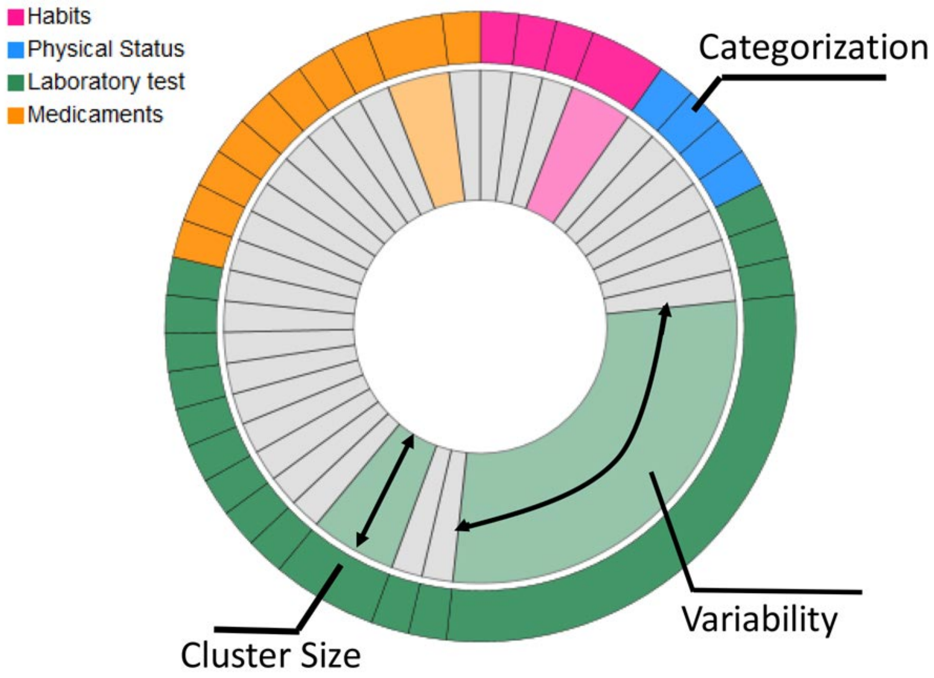


Figure 4.7: Each subspace cluster is illustrated as a two-ring donut plot. Sectors are representatives of features and are colored based on the categorization of the data. The outer ring shows the categories of features. Thus, the analyst gets insights about which features from which category are (not) involved. The length of sectors represents the variability of the corresponding features.

outermost ring stands for feature categorizations. The non-involved features in the selected subspace cluster are on grayscale, while the involved features are in the color scale of their assigned color. Therefore, darker colors represent larger values and, in contrast, brighter colors depict smaller values. The missing values are differentiable as they are in black.

In fact, each ring in the donut chart represents a participant. Due to having a lot of members, the detailed information of each patient is not visually tractable. Thus, S-ADVIsED allows the user to have access to the table of individuals' features by selecting the ring of the corresponding participant (T3).

S-ADVISIsED also provides an optional sorting strategy based on the feature that has the greatest variance. In this way, the participants who have greater variance will have a bigger radius.

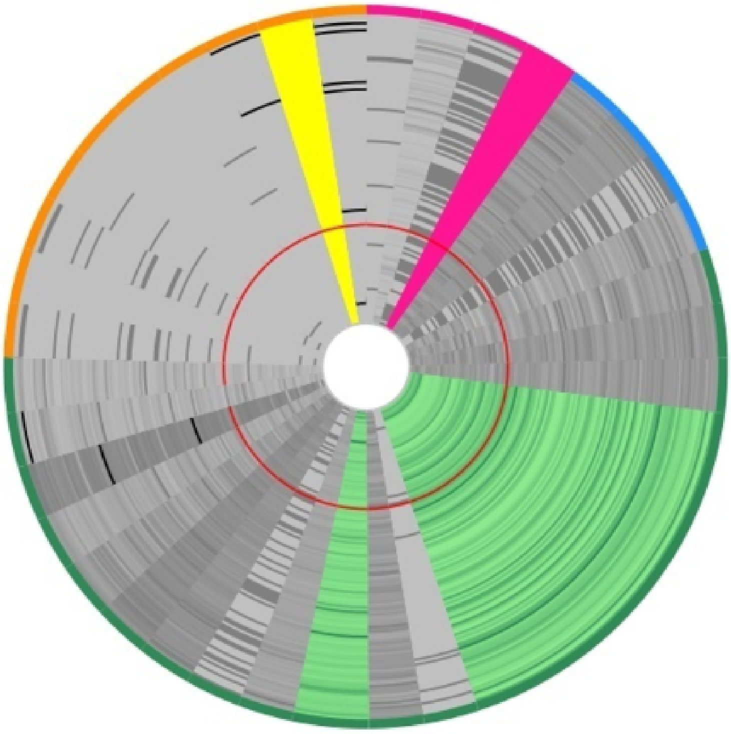


Figure 4.8: Donut heatmap shows detailed information on the values of a selected subspace cluster. For the involved features, the color range is based on the color categorization of features, while non-involved features are in grayscale. The missing values are colored black.

**Mosaic Plot**: Mosaic plots are embedded to show the relationship between categorical features, see Fig. 4.9. The user is enabled to make an on-demand selection of a categorical feature and see its relation to another selected target feature, e.g. sex and fatty liver (T5).

**Scatterplot matrix (SPLOM):** A discretized SPLOM with embedded histograms shows the relationship between the numeric features of a selected cluster. The discretization is based on a target feature, e.g. fatty liver, which is distinguishable via colors. For example, in Fig. 4.12 (b1-b2), blue points are representative of negative fatty liver participants, whereas the orange ones are participants with positive

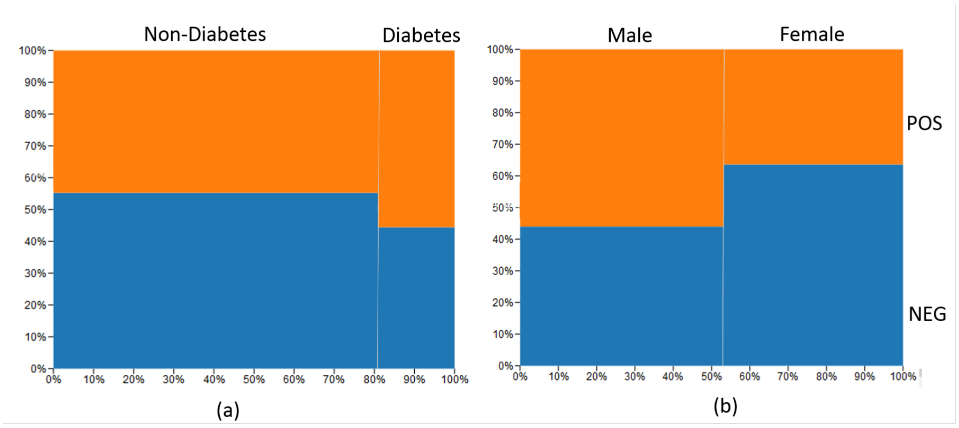


Figure 4.9: Mosaic plot shows the relationship between categorical and a selected target feature. The proportion of participants regarding diabetes and fatty liver (a) and the relationship between gender and fatty liver in the oldest subspace population (b).

fatty liver. The histograms on the main diagonal show the distribution of each feature (T6).

**Stacked error bars:** Stacked error bars represent the relationship between a selected numeric feature, e.g. BMI, and a categorical selected target feature, e.g. fatty liver. As presented in Figure. 4.10, positive and negative values are normalized to form the average value of the corresponding cluster. All bars are sorted based on the average value, while the error lines depict the standard deviation of the selected feature (T5).

As shown in Figure 4.10, the sub-cohort number 10 with 50 participants had the highest proportion of positive fatty liver participants. This sub-cohort has the highest average BMI values, i.e. 34.2±.06, as the participants mainly suffer from obesity. Moreover, this is the oldest sub-cohort with 58±9 years old participants. The participants in sub-cohort number 10 also have the highest thyroid volume (21.7±6), the highest average amount of glucose (6.3±2.3) and the lowest level of TSH, which is a hormone that controls the thyroid gland activity.

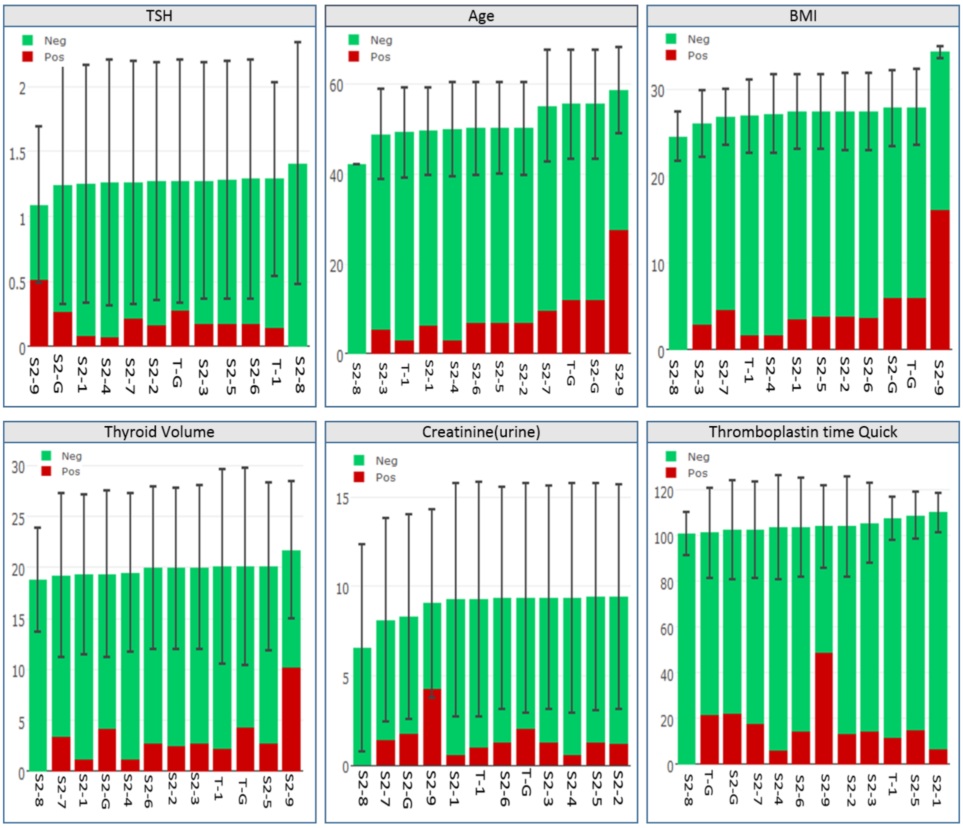


Figure 4.10: The stacked error bar shows the mean value of different features for all subspace clusters. The error line depicts the standard deviations for the selected feature.

### Cluster description and validation of findings

Firstly, subspace clusters should be described as a form of rules. To achieve this, they need to be defined as intervals within a hyper-rectangular shape, e.g. a sub-cohort with $BMI > 30.5\, kg/m^2 \wedge TSH \leq 1.5\, mU/l$, 52 % suffer from goiter.

Secondly, based on the discussions with Henry Völzke and Till Ittermann, epidemiology experts do not easily trust and accept the results achieved by the data mining techniques, i.e. subspace clustering. Thus, the results need to be verified somehow.

Usually, one approach to verify the subspace clustering results is replication. However, it means if we have the results of sub-cohort in dataset/ cohort $DS_A$, it is necessary to have access to another independent dataset/

cohort $DS_B$. Sub-cohort $SC_A$ is validated, if it can be reproduced as $SC_B$ in an independent cohort $DS_B$.

S-ADVIsED does the cluster description and validation in one step (T8 and T9). It allows the expert to simultaneously adjust the shape of the subspace cluster in specific intervals and checks if the cluster is replicated in another cohort. According to the meeting with Henry Völzke and Till Ittermann, several measures should be considered to verify the reproducibility of sub-cohort $SC_A$ and sub-cohort$SC_B$ [68]:

- **Involved features:** $SC_A$ and $SC_B$ should have the same involved features in a subspace-cluster.
- **Distribution:** Both $SC_A$ and $SC_B$ should have the same distribution and variance regarding the involved features and the target feature.
- **Sub-cohort size:** $SC_A$ and $SC_B$ should have a relatively close/same number of participants, e.g. 5% of individuals in each dataset.
- **Deviation:** $SC_A$ and $SC_B$ deviate from the global mean (whole cohort) in the same way regarding a distinct feature.

To apply cluster description and verify replication of a sub-cohort, S-ADVIsED proposes the following steps (Sections. 8 and 9):

1. **Subspace cluster selection:** Activating the replication task (4.6(c)) and selection of $SC_A$ subspace cluster from the global view (4.6 (d)) based on interestingness, based on its similarity to the other clusters, involved features and the number of involved individuals.
2. **Visual classification:** Multiple sub-cohort candidates can be defined by the user. For each candidate the following steps should be provided:
   a. Concatenating the members of $DS_A$ and $DS_B$ and visualizing them in a scatter plot matrix (SPLOM) with embedded histograms distinguishable by colors. For example, Figure 4.12 shows the members of $DS_B$ in green, while the members of $DS_A$ are colored based on the target feature which is fatty liver (negative as blue and positive as orange).

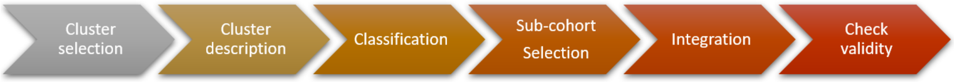


Figure 4.11: The overall steps for sub-cohort discovery and validation of subspace clustering results.

   b. The user is allowed to do the cluster description. To achieve this, the user can define desired intervals by drawing rectangles within each pair of features. As presented in Figure 4.12(a1-a3), the linking and brushing technique lets the user see the selected members within the rectangle in other pairs of features. Moreover, the diagonal histogram shows the distribution of each feature.

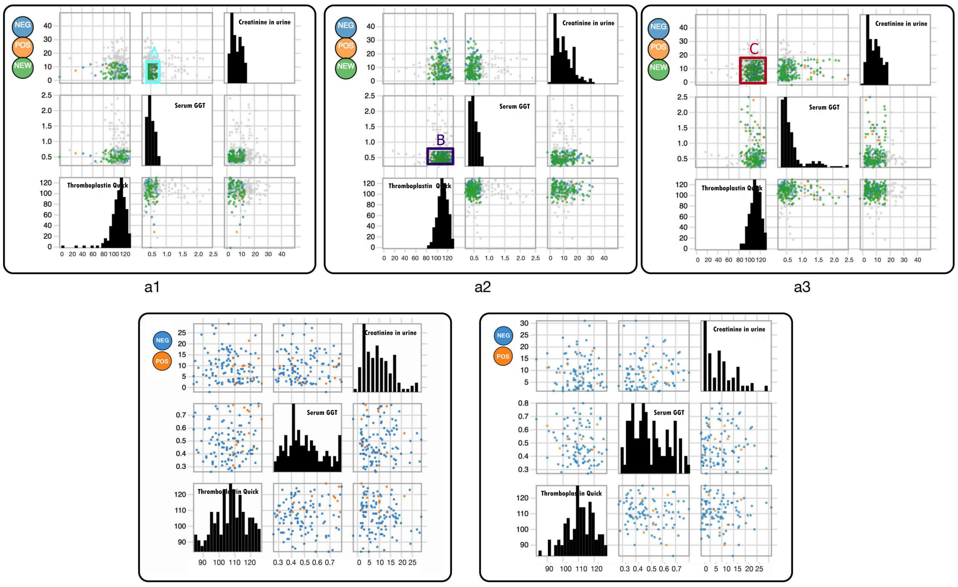
   c. The labels of $DS_B$ members are predicted by inheriting the label of the closest $DS_A$ neighbor withing the drawn rectangle, i.e. 1-nearest neighbor. There is an assumption that it is more likely that individuals closer to each other have the same label.

3. **Sub-cohort selection:** In this step, the ROC curve shows the relationship between the true positive rate (tpr or sensitivity) and the false positive rate (fpr or 1-specificity) of the candidate sub-cohorts. Then, the expert can pick one of the candidate sub-cohorts based on the tpr and fpr values. For example, if the expert is interested in a sub-cohort with more fatty liver members, (s)he may pick up the sub-cohort from the ROC plot with higher tpr and lower fpr values (Fig. 4.13).

4. **Integration:** In the last step, the $SC_A$ (derived from the results of subspace clustering) and $SC_B$ (derived from the replication process) are shown in the global view for further assessments described in Section 4.4.2 using the exploration features of the system Section 4.4.2.

## 4.5 Subpopulation Discovery in Longitudinal Data

To analyze the sub-cohort in longitudinal data, we designed mock-ups to find the sub-cohort in different waves of study using bi-clustering techniques and investigating the sub-cohorts' transitions during the time.



a1 a2 a3

b1 b2

Figure 4.12: (a1-a3) illustrate the selected intervals. Fig. (b1) is the distribution of the newly generated subpopulation labeled T-1. Fig. (b2) illustrates the scatter matrix of the S2-1 subpopulation which is transformed to the new intervals.

### 4.5.1 Task analysis

The tasks consists of four functionalities, see Fig. 4.14:

1. **Identifying sub-cohorts**: The first step is to identify sub-cohorts. To achieve this, we intend to apply bi-clustering in one selected wave of the study. It finds a subset of features and participants who are highly correlated to each other.

2. **Inspection of sub-cohorts**: Then the identified sub-cohorts should be analyzed. The comparison could be related to the correlation between involved features and also the change of correlation between features should be observed during different time points.

   The distribution of involved features should be visualized to give expert insight into the results. Additionally, the analyst should see the transition of sub-cohorts over time. It helps to understand whether a cluster is changed.

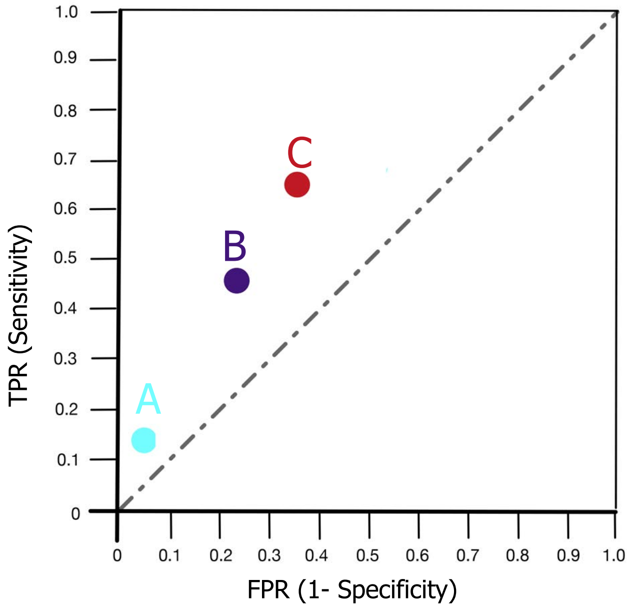
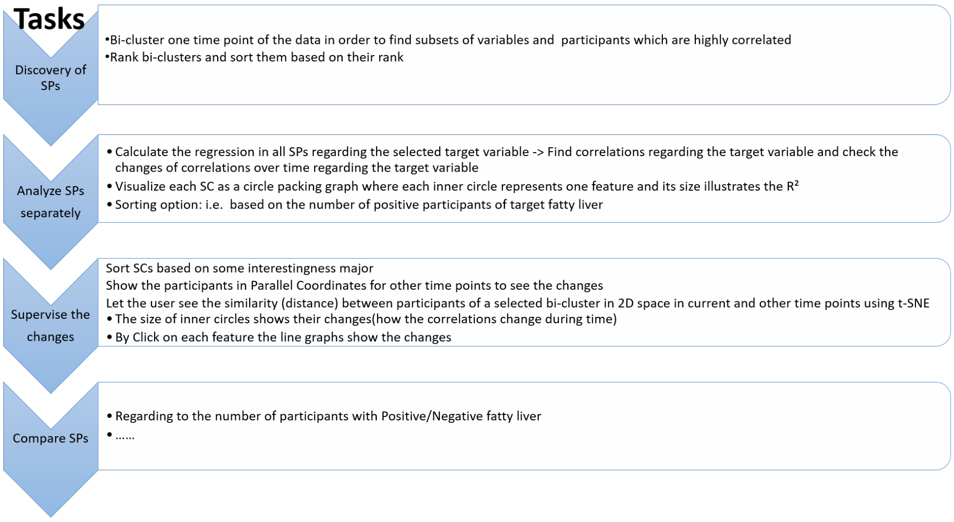


Figure 4.13: Each point in the ROC curve depicts the TPR and FPR ratio for the selected rectangle. Each point corresponds to a drawn rectangle in Fig. 4.12 (a1-a3).

### 4.5.2 Mock-ups

Figures 4.15-4.17 shows the mock-ups of different views of the intended system. In the first step, the user shall select a specific time point of the study from the left-side menu as starting point and as input to the bi-clustering algorithm. Then, the type of bi-clustering algorithm shall be selected. It is supposed to apply column-based bi-clustering algorithms [69]. Then, start the discovery of finding bi-clusters.

As presented in Fig 4.15, the SHiP-1 dataset was selected as a starting point and the *Gabi* algorithm was selected as a bi-clustering method [69]. The features are categorized the same as the S-ADVIsED and a specific color is assigned to each category. The list of sub-cohorts is sorted based on a ranking criterion such as the number of positive participants for the outcome feature. The discovered bi-clusters are listed in the top panel. Each bi-cluster is presented as a packed circle plot, as each inner circle is related and to an involved feature and colored based on the category of features. Each involved feature is ranked based on the $R^2$ value encoded as the size of the circle. Thus, the size of each feature (the inner circle) illustrates how



- Bi-cluster one time point of the data in order to find subsets of variables and participants which are highly correlated
- Rank bi-clusters and sort them based on their rank

- Calculate the regression in all SPs regarding the selected target variable -> Find correlations regarding the target variable and check the changes of correlations over time regarding the target variable
- Visualize each SC as a circle packing graph where each inner circle represents one feature and its size illustrates the $R^2$
- Sorting option: i.e. based on the number of positive participants of target fatty liver

Sort SCs based on some interestingness major
Show the participants in Parallel Coordinates for other time points to see the changes
Let the user see the similarity (distance) between participants of a selected bi-cluster in 2D space in current and other time points using t-SNE
- The size of inner circles shows their changes(how the correlations change during time)
- By Click on each feature the line graphs show the changes

- Regarding to the number of participants with Positive/Negative fatty liver
- ……

Figure 4.14: Workflow of the prototyping of the intended system for sub-cohort discovery in longitudinal cohort studies.

it is associated with the target feature. To calculate $R^2$ values, the regression of all involved features is calculated w.r.t. the target features such as the fatty liver. The $R^2$ shows how each feature influences the target feature.

In the next step, the user can select one of the bi-clusters from the top panel and observe how the correlation of features regarding the target feature is changed during different time points. For example, in Fig. 4.15 the user selected a bi-cluster with four involved features consisting of age, smoking, cholesterol, and creatinine status. The $R^2$ tab of the bottom panel shows how the dependency of each feature regarding the target feature is changed.

Next, as shown in Fig. 4.16, to check the transition of sub-cohorts (bi-clusters) over time, the involved participants are visualized in 2D space by a dimensionality reduction technique like MDS. The participants are shape-coded w.r.t. the target feature, e.g. the circle represents positive and rectangle shows negative participants. By using the timeline located at the bottom of the 2D projection plot the user can see how the clusters evolve during the time or change the class for the target feature.

Next, for further analysis of the bi-cluster by switching to the distribution tab the user can see the distribution of the involved feature for the selected bi-cluster in a SPLOM. The user is allowed to see how the distribution of involved features is changing in different time points by applying the timeline located at the bottom part of the panel.

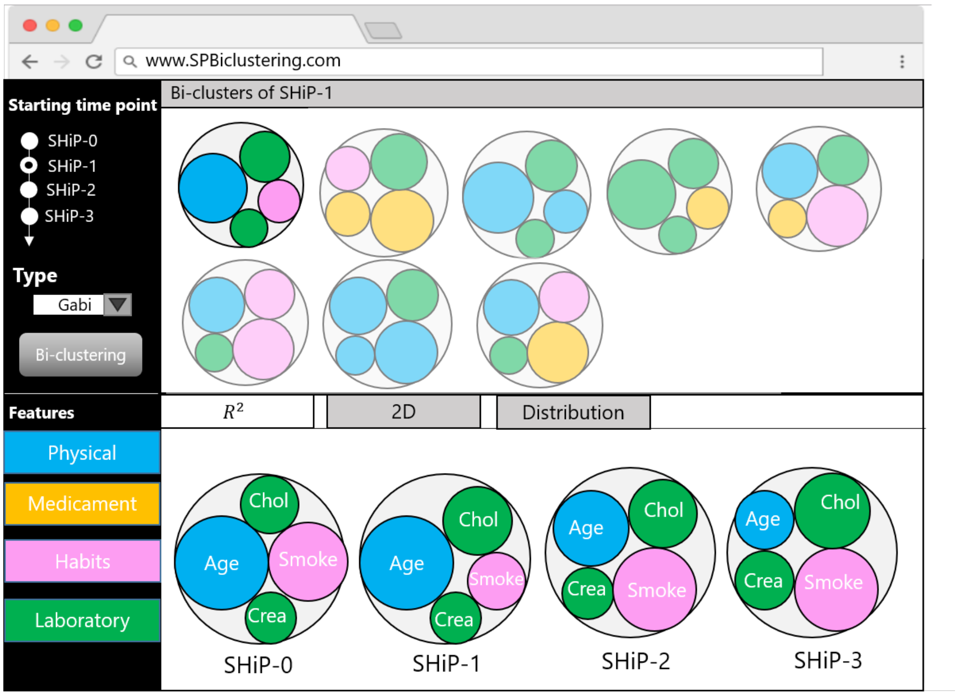


Figure 4.15: The prototype consists of two main panels. The top panel shows the identified bi-clusters (e.g. SHiP-1) in the form of packed circles. From the bottom panel, the analyst is allowed to analyzed a selected bi-cluster in different time points. The $R^2$ tab shows the involved ranked features (based on $R^2$ values w.r.t. the target feature) of the involved bi-clusters in different waves of the study.

## 4.6 Summary and Conclusion

In this chapter, S-ADVIsED as a system for sub-cohort discovery on epidemiological data is described. The requirements were have collected based on on-site visits at the Epidemiology Department of the Medical University Greifswald. Then S-ADVIsED as a web-based system was designed to enable the expert users to explore, extract and validate the sub-cohorts from the clustering result.

The focus of S-ADVIsED is on cross-sectional data. Thus, the mock-ups of a new system are described in this chapter for the analysis of longitu-

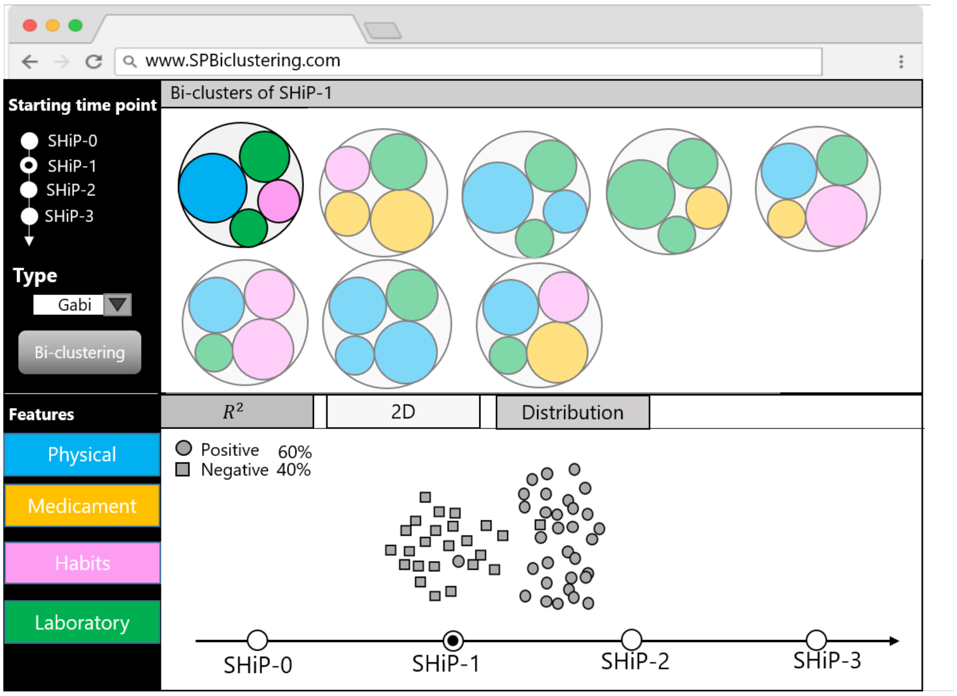


Figure 4.16: The 2D tab plots the distance between the members of the selected bi-cluster in the 2D space by a dimensionality reduction technique. The class of the target feature is distinguishable via the shape of points. The switching between waves of the study using the timeline shows the transition of a bi-cluster in previous and next waves.

dinal data. The aim of the system is to compare the different sub-cohorts identified by the bi-clustering algorithm over time.

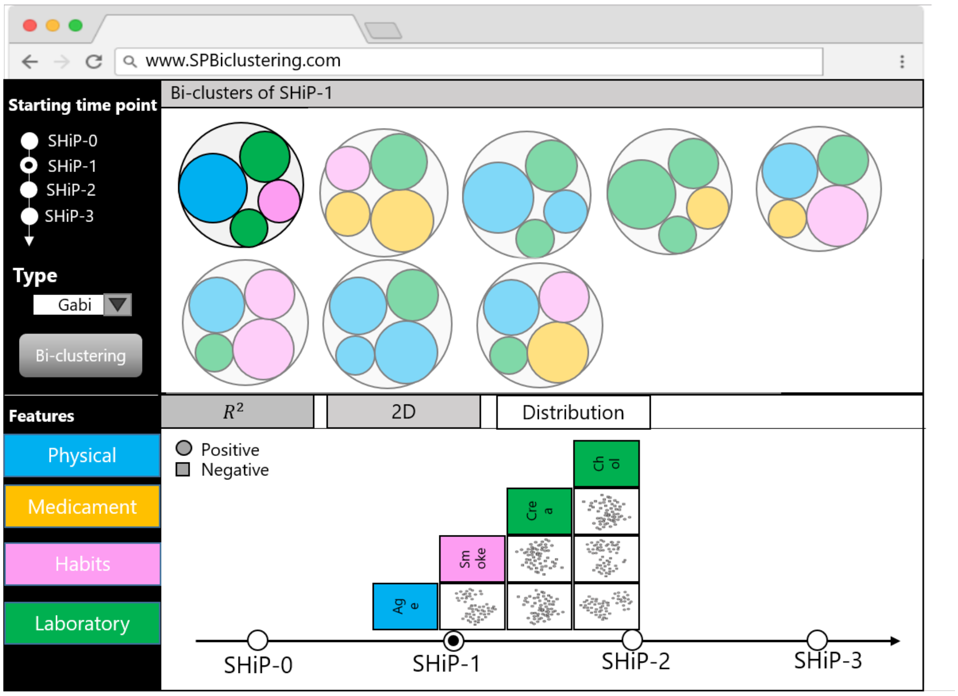
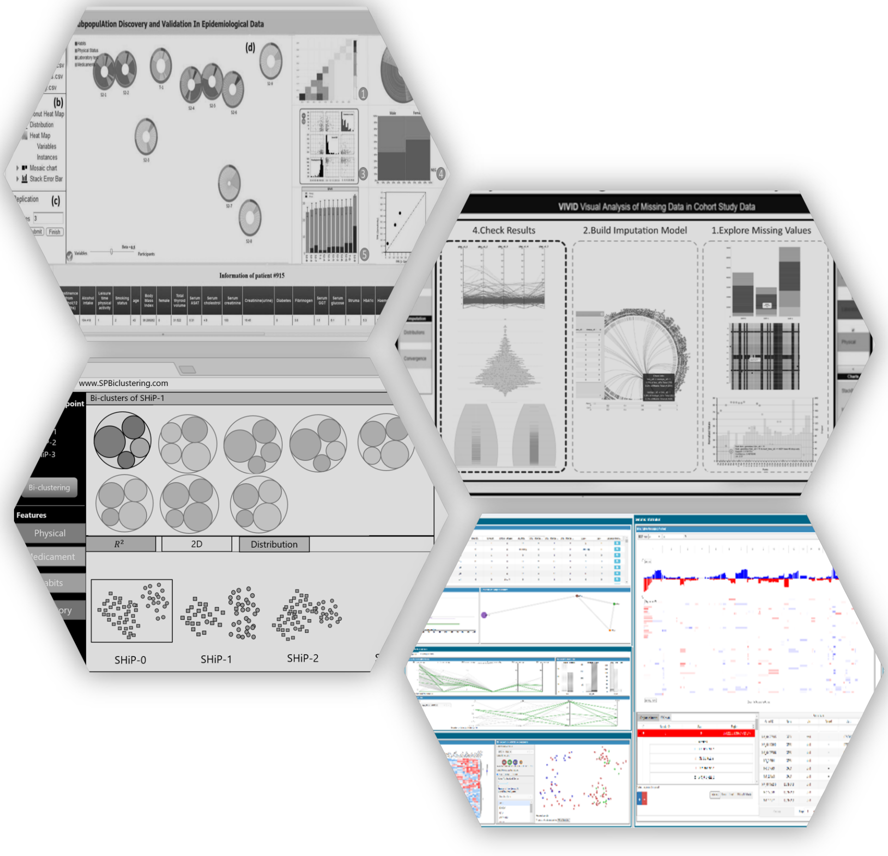


Figure 4.17: The distribution tab visualizes the distribution of involved features of a selected bi-cluster using SPLOM. The analyst gets insights into the changing of the distribution of a sub-cohort in different time points w.t.r. each involved feature by switching between different time points.

# 5

# Sub-cohort Discovery of Multi-Omics Data



This chapter is based on

- A Visual Analytics Approach for Patient Stratification and Biomarker Discovery. Shiva Alemzadeh, Florian Kromp, Bernhard Preim, Sabine Taschner-Mandl, Katja Bühler. Proc. of Eurographics Workshop on Visual Computing for Biology and Medicine (EG VCBM), pp. 91-96, 2019.

In modern diagnostics, grouping individual patients into risk groups help physicians to make better decisions for the treatment of patients of each group. Precision medicine, personalized treatment, is not a new topic, since, in many studies like the NCI MATCH trial or the INFORM study [70], the precise diagnostics that help to find out effective individualized treatment based on molecular abnormalities was already introduced. One of the main aims of precision medicine is to identify subgroups of patients that had a better response to a specific treatment in comparison to the average population. For this purpose, diagnostic and prognostic biomarkers should be discovered [71].

Nowadays, the traditional histopathological techniques, such as microscopic examination of tissue for identifying a disease, are replaced by the molecular diagnostics approaches which refer to a set of techniques for analyzing the biological markers in the genome code (i.e. how each person's genetic code expresses their genes as proteins), as it offers better accuracy and sensitivity for biomarker discovery [72]. Understanding the tumor-specific molecular changes in the genome-wide search, i.e. OMICS, leads to a better specification of tumorigenesis as well as progression and relapse or treatment failure.

Identifying the groups of tumor types, risk groups and the prediction of survival asks for information of a DNA-based analysis (genomics) and gene expression analysis (messenger ribonucleic acid mRNA). Moreover, involving the clinical and minimal residual disease (MRD) into the analysis leads to more precise results. MRD refers to the number of cancer cells in the bone marrow sample which are leftover in the patient during and after the treatment procedure. A leukemia the patient is considered in remission when there is no MRD or sign of the disease in the body after treatment [73].

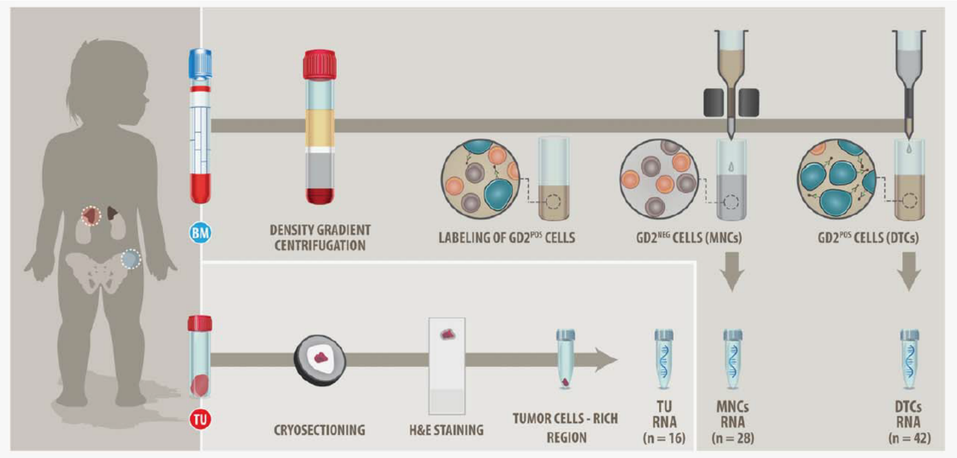


Figure 5.1: Neuroblastoma originating from adrenal glands and involving other parts of the body. The process of mRNA and bone marrow sampling is shown [80].

This approach involving information of patients from multiple sources is specifically effective for rare cancers [74]. Any type of tumor in any age that involves a small portion of the population is considered as rare cancer [75]. Surprisingly, rare cancers, unlike their name suggests, are not rare, as annually 541,000 people in Europe are diagnosed with rare cancers, which stands for 20% of all cancer patients. Unfortunately, due to the unknown characteristics of rare cancers and the lack of large studies, the five-year survival rate for these patients is considerably lower than for most of the common cancers,i.e. 47% vs. 65% [76][77].
In this chapter, the data of neuroblastoma cancer patients as a type of rare childhood cancer was analyzed. Neuroblastoma is a rare childhood cancer with a portion of 7-10% which originates from the adrenal glands (Figure 5.1). Actually, it is the most common cancer type in infants younger than one-year-old, which accounts for 15% of all cancer-related deaths in children [78]. Despite intensive multimodal treatment, the overall survival of high-risk patients is still below 50% due to therapy refractory or relapsing disease [79].

Currently, to identify high-risk and low-risk groups of relapse or death using different kinds of data of these patients, cancer research is independently using various tools. There is a need for a joint exploration of these data which leads to the identification of distinct sub-cohorts. Thus, DiscoVA as a visual analytics multiple coordinated views system for patient stratification and biomarker discovery is introduced to cover this gap.

The main contributions include:

- Design and development of DiscoVA for joint exploration and sub-cohort discovery using clinical, MRD, and wide-scaled multi-omics data. The DiscoVA consists of the following components:
  - Integrated inspection of patients data from multiple sources consisting of clinical, MRD, genomic somatic copy number aberrations (CNA) / (DNA) and mRNA data
  - Iterative hierarchical sub-cohort discovery by employing heterogeneous data
  - Functionalities to compare sub-cohorts
- Evaluating the utility of DiscoVA by carrying out an evaluation with biomedical experts to validate a hypothesis based on the previous studies and rating the system by providing a survey form.

## 5.1 Related Works of VA

Visual analytics using OMICs data, and especially for precision medicine, is an understudied subject. The main aim of this research is to help to make decisions for the treatment of patients by considering their unique information. There are some recent works in this field which are described in the following.

Marai et al. [81], developed a multiple coordinated view system for the exploration of heterogeneous neck cancer patients' data. They introduced a novel nomogram plot to describe the probability of the patient's survival using a K Nearest Neighbor approach. The system consists of several visualization techniques including Kaplan-Meier estimator, mosaic plots to describe the cohort characteristics along Kiviat plots to show each individual's features. Figure 5.2 shows an overall schema of the system. Although the system provides an appropriate solution for patients' exploration in

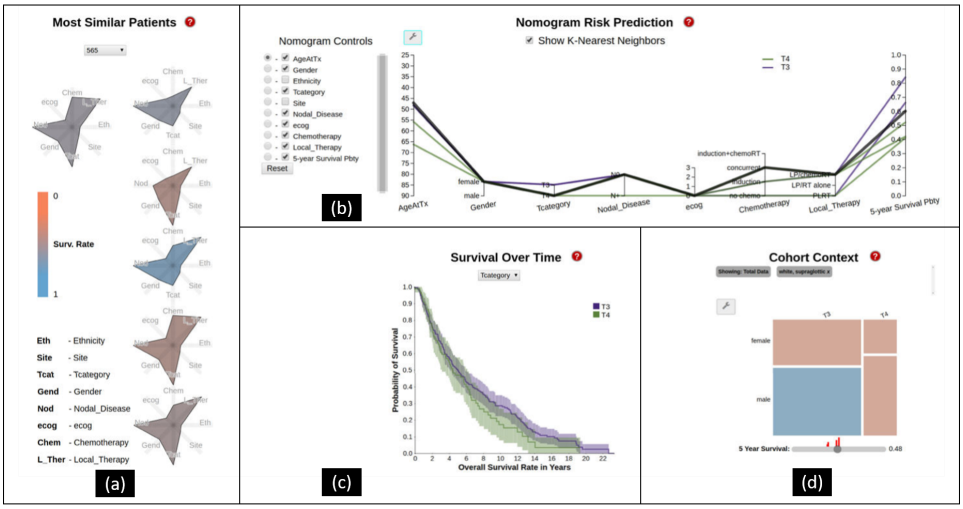


Figure 5.2: (a) shows the patients' features along with the most five similar patients using Kiviat plots. The color coding represents the patient survival probability. (b) The introduced nomogram describe depicts variation in treatment. (c) The Kaplan-Meier plot shows the survival of cohorts over time. (d) Mosaic plot represents the relationship between sex and tumor stage [81].

precision medicine, there is a lack of using the genome and mRNA data of patients.

[82] developed a web-based system called Hitwalker2 for analyzing the patients' gene profile using a ranking technique for each individual. The main focus of Hitwalker2 is the genomic data of patients.

[83] developed StratomeX as a cluster browser to explore the correlation of patient stratification using OMICS data. As shown in Figure. 5.3, it uses heatmaps, pathway maps along with Kaplan-Meier plots to show the relationship between the patients' groups. StratomeX was extended by [84]. In this work, StratomeX was integrated with an inspection system that allows the comparison of groups of patients using their heterogeneous clinical, genomic, and molecular profile data (Figure 5.3).

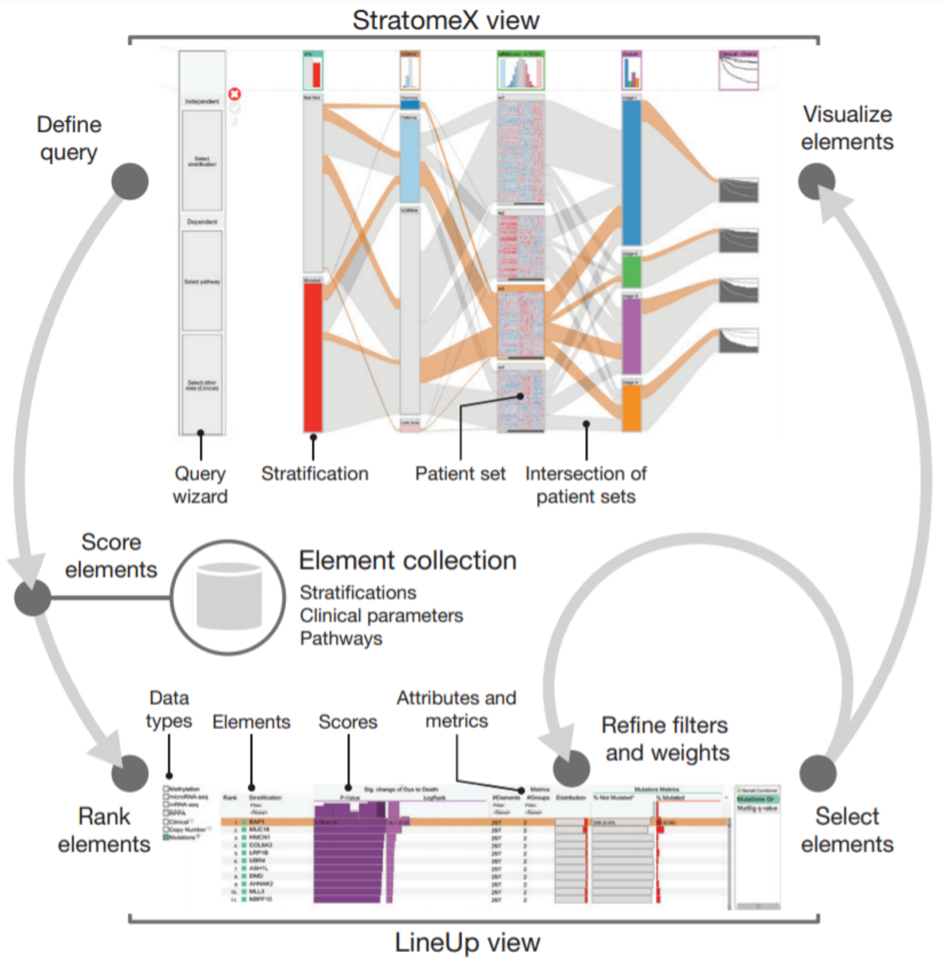


Figure 5.3: Extended StratomeX for exploration of patient strafications. The top view shows the stratifications of patient sets. The bottom view shows the results of queries by listing the ranked elements [84].

DiscoVA is different from previous works from several perspectives. It allows the joint navigation of heterogeneous data and enables the expert to make complex nested visual queries on a combination of multi-omics and clinical data.

## 5.2 Background and Data

The most common cancers are well characterized at the molecular level. Nowadays, the main trend is to integrate clinical data with OMICS data collected from different tissue sites like primary tumor or metastasis in different time points, i.e. from diagnosis, during chemotherapy, and relapse.

The genetic characteristics of rare cancers that are related to the relapse of the disease are less known. The reason is that the number of patients is not enough to pursue studies on these patients. Moreover, their clinical and biological features are highly heterogeneous. As an example, in adulthood cancers, common solid tumors are driven by recurrent somatic mutations, which is less likely in childhood cancers such as neuroblastoma [85].

Metastatic cancer has a high risk of relapse which is mostly due to small remaining parts that were not destroyed in the initial treatment. Thus, predicting the chance of relapse/death by identifying risk groups is one of the focuses of cancer research. Childhood cancers are usually characterized by:

- CNAs which are segmental chromosomal aberrations resulting in large-scale gains and losses, i.e. certain genes occur twice and others are missing as a result of copying errors.

- gene amplification and smaller intragenic deletions, as well as translocations.

- genetic alterations as predictive factors for a dismal clinical outcome. Or they have been found enriched in relapse tumors or metastases, e.g. *MYCN*-amplification, the presence of certain segmental chromosomal aberrations, amplification or mutations in the ALK gene, or the activation of telomere maintenance mechanisms (ATRX deletions, hTERT overexpression). Thus, diagnostic and data exploration workflow requirements are strongly dependent on the genomic alterations relevant to the tumor type of interest.

- gene expression classifiers are able to discriminate quite robustly high- from low-risk patients and the integration of clinical and molecular classifiers can substantially improve the risk stratification [86] [87].

---

[0] a change in the DNA sequence of a somatic cell that refers to any biological cell which forms the body of an organism

As more than 95% of neuroblastoma patients with high-risk metastatic disease present with tumor cells infiltrating the bone marrow, the highly sensitive detection of these tumor cells by automated immunofluorescence (a technique to determine the location of an antigen/ antibody) imaging approach is currently state of the art to evaluate whether patients respond to treatment, and to monitor disease (MRD) in current neuroblastoma clinical trials [88]. The dynamics of bone marrow clearance are currently used to identify patients responding to treatment at an early time point in ongoing studies.

In this work, data from the high-risk group of neuroblastoma patients from the Children's Cancer Research Institute in Austria was used. Sabine Taschner-Mandle, Florian Kromp, and Fikret Rifatbecovic provided the neuroblastoma patients' data for the analysis. The data consists of information from four different sources:

1. **Minimal Residual Disease (MRD) data:** In this study, MRD stands for the number of cancer cells in bone marrow samples [89]. It might be a prognosis marker for the prediction of relapse and a measure to show the effectiveness of the therapy for an individual patient. The MRD data of bone marrow samples were evaluated for 559 patients by automated immunofluorescence microscopy plus FISH (AIPF) on cytospin preparations at 8-time points of the disease course.

2. **Clinical data:** The clinical data consists of features such as blood and sample values, metastasis, life and relapse status, and other features that may be correlated with disease relapse/prognosis. The clinical information in this study consists of 62 features of 559 patients.

3. **DNA copy number aberration (CNAs) data:** CNA data shows the genomic interval abnormalities, i.e. gained and lost regions that were determined semi-automatically, and due to over-segmentation they were later manually curated. In this study, the CNA information was available for 20 patients.

4. **mRNA expression data** RNA sequencing data have been generated from primary tumor tissue and bone marrow disseminated tumor cells and were collected at four different time points. For analysis, mRNA data were available for 20 patients on 20202 genes.

## 5.3 Limitations

Integration, identification, and classification of rare cancer patients' data including neuroblastoma using classical data mining techniques is a non-trivial task due to the following reasons:

- Integration of these heterogeneous data is not always possible due to the high dimensionality and different structure of data, e.g. different scales. For instance, in this research, the clinical data scale and type are not comparable with DNA and mRNA data which consists of information of thousands of genes.
- Due to an insufficient number of patients in each group, it is not always possible to guarantee the statistical significance of the learned models.
- As for biomedical experts, it is necessary to validate the findings of data mining techniques, usually, there is a need to use another dependent dataset or a part of the main dataset for validation, but because of the lack of adequate studies, it is not feasible.
- The dataset of patients is usually sparse and missing data is still an issue in these data, as not all information is available for every patient. As an example, the RNA and DNA study has only been performed on a subset of patients, as often data collection is expensive or the biopsy was not enough for both DNA and mRNA analysis. Most of the automated or semi-automated methods need the availability of all or a major part of the data.
- Because of the lack of a visual analytics tool, joint exploration and sub-cohort discovery using all heterogeneous OMICs and clinical data of patients is not achievable.

These facts hamper the exclusive application of classical machine learning methods and/or deep learning for the identification of patient sub-cohorts. In the absence of sufficiently powered unbiased approaches, biomedical researchers need convincing evidence for the robustness and accuracy of their data mining results. In order to close this gap, we propose to build on expert knowledge in a semi-automated manner.

## 5.4 DiscoVA system

The following sections describe the different aspects of the design and implementation of DiscoVA system.

### 5.4.1 System Design and Task Analysis

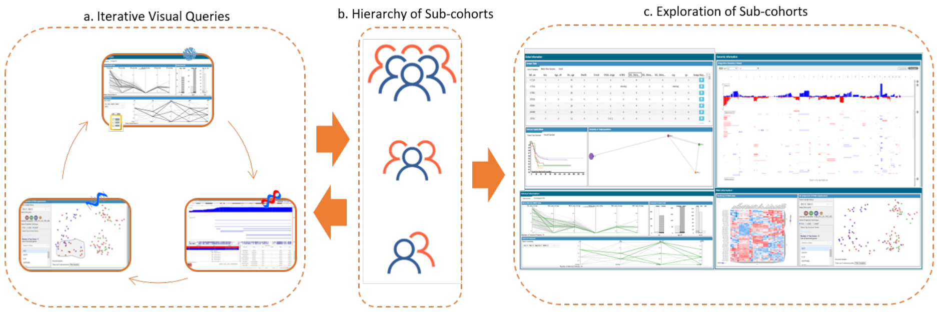

a. Iterative Visual Queries
b. Hierarchy of Sub-cohorts
c. Exploration of Sub-cohorts

Figure 5.4: The overall workflow of DiscoVA consists of (a) the user performing iterative queries on the different sources of data to identify sub-cohorts. (b) After identifying the sub-cohorts, the expert is allowed to save the hierarchy of sub-cohorts in the database. (c) The stored sub-cohorts can be retrieved from the database for further investigation.

The design decisions of DiscoVA were made in close cooperation with the data scientists and biology experts from Saint Anna Children's Hospital. The system development consisted of three major phases. Several meetings were held within each development phase to optimize the system design and specifications:

1. System design and analysis of requirements: In the first phase, biology experts introduced visualization techniques commonly used in their field, such as Kaplan-Meier plots, mRNA data cluster heatmap, or IGV for genomic data analysis. Then, the mock-ups were designed to show the structure and features of the system to the biology experts. Moreover, they were helpful to define the strategies for future development.

2. Implementation of the software: In the second phase of the development, the implementation of the system's components were prioritized based on the current limitations, e.g. availability of the data.

For example, the clinical and MRD data were available from the beginning of the development. Thus, it was decided to start with the implementation of the components for clinical and MRD data.

3. Follow-up: In the third phase, the ultimate design and configuration of all components have been addressed.

According to the design decisions, the system requirements were established to allow distinct sub-cohorts to be discovered based on a joint exploration of the multi-omics data along with clinical data. The tasks consist of:

- T1. Joint exploration of heterogeneous patient data.
  - (a) **Clinical and MRD data:** Visualization of clinical and MRD data features should be provided to give insight to the expert about the distribution and correlation between features. Moreover, the expert should be allowed to visually select a sub-cohort (group) of patients based on these features.
  - (b) **Genomic data:** The genomic/CNA intervals should be compared to allow the identification of sub-cohorts with genomics aberrations by exploration and digging into different chromosome intervals. The expert should be enabled to select a set of regions of interest and make queries by getting the union, intersecting, and inverting the list of patients in regions of interest.
  - (c) **mRNA expression data:** There is a need to show the groups of samples which share common transcript levels.
- T2. **Iterative sub-cohort discovery:** To allow hypothesis generation, distinct sub-cohorts should be identified by utilizing various aspects of the data. The expert should be enabled to interactively make queries on heterogeneous data, i.e. MRD, clinical, CNA/DNA, and mRNA.
- T3. **Inspection of individuals:** Analysis of features of individual patients should be provided in order to have a more detailed view on the members of the cohort, e.g. analysis of different sample tissues in different time points for the same patient.

T4. **Comparison of sub-cohorts:** The overview of sub-cohorts should be provided in order to display how similar they are to each other regarding overlapped patients. Moreover, they should be compared regarding the survival rate outcome. It gives the expert insight on whether the patients of a sub-cohort had a poor prognosis or how successful they were in the treatment.

### 5.4.2 Components

As a part of this thesis, the DiscoVA system as an interactive multiple coordinated system was developed to cover the requirements discussed in Sect. 5.4.1. The whole system consists of four main parts comprising of:

- Overview section
- Clinical and MRD section
- Genome info section
- mRNA info section

The corresponding dataset of each section is loaded at the system startup. Generally, DiscoVA's user interface (UI) has been designed to assist biology experts in investigating and identifying sub-cohorts in a specific section, e.g. clinical data, and highlighting the sub-cohorts in other sections. Then, the expert can continue and narrow the query with other types of data, e.g. mRNA data. In each step, the user is allowed to save sub-cohorts for later reference in a hierarchical manner. As shown in Figure 5.5, the hierarchy of sub-cohorts is accessible via the left side menu. The application follows the color-coding as below.

| **Color** | **Visualization** |
|---|---|
| Gray | Whole cohort |
| Black | Brushed sub-cohort |
| Red | Overlap information |
| Others | Saved sub-cohorts |

A unique color is assigned to any sub-cohort, where the child of each sub-cohort gets a brighter color of its parent sub-cohort. The color of each sub-cohort is calculated as Eq. 5.1, where the $\beta$ value corresponds to the level of brightness which is between 0 and 1.

$$\begin{aligned} R_j &= min(255, R_i + (255 * \beta)) \\ G_j &= min(255, G_i + (255 * \beta)) \\ B_j &= min(255, B_i + (255 * \beta))) \end{aligned} \tag{5.1}$$

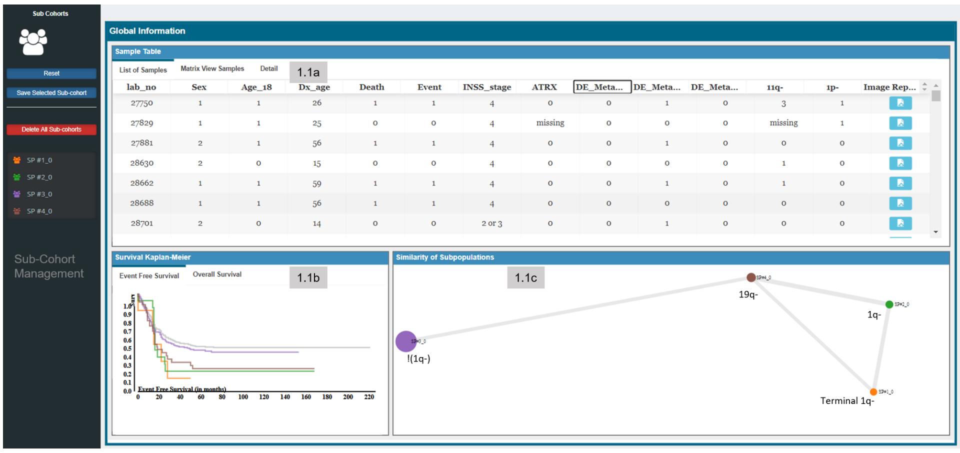
In the following, we describe each section of the DiscoVA system in more detail.

**Overview section**

This section gives information on identified sub-cohorts and members of each sub-cohort (T4).

a. **Individuals' information:** This component summarizes individuals of sub-cohorts including their genomic and clinical information. This view is provided in three tabs:

   - The first tab shows the clinical information of a set of selected features for patients of the currently selected sub-cohort in a table view. Figure 5.7 shows the list of male patients.
   - The second tab summarizes the genomic aberrations of the current sub-cohort using simplified circos plots introduced in this work. Thus, the expert can perceive abnormalities in different chromosome levels of the current sub-cohort at a glance. As shown in Figure 5.6(1.2a), each circle represents the CNA/DNA information in a sample. The number of sectors is equal to the number of chromosomes (i.e. 22 sectors) and the length of the corresponding chromosome. The circos plots are colored based on the conventions in biology, i.e. red represents losses and blue represents gains. If a sample has at least a gain or deletion in a chromosome the corresponding sector in circos plot gets blue and red, respectively. In addition, if there are both gain and deletion intervals in a chromosome, it is shown in yellow. The neutral sectors are gray.
   - The third tab shows the detailed information of the selected patients including the table of clinical information and the list of cir-



| lab_no | Sex | Age_18 | Dx_age | Death | Event | INSS_stage | ATRX | DE_Meta... | DE_Meta... | DE_Meta... | 11q- | 1p- | Image Rep... |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 27750 | 1 | 1 | 26 | 1 | 1 | 4 | 0 | 0 | 1 | 0 | 3 | 1 | |
| 27829 | 1 | 1 | 25 | 0 | 0 | 4 | missing | 0 | 0 | 0 | missing | 1 | |
| 27881 | 2 | 1 | 56 | 1 | 1 | 4 | 0 | 0 | 1 | 0 | 0 | 0 | |
| 28630 | 2 | 0 | 15 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 1 | 0 | |
| 28662 | 1 | 1 | 59 | 1 | 1 | 4 | 0 | 0 | 1 | 0 | 1 | 0 | |
| 28688 | 1 | 1 | 56 | 1 | 1 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 28701 | 2 | 0 | 14 | 0 | 0 | 2 or 3 | 0 | 0 | 1 | 0 | 0 | 0 | |

Figure 5.5: The first tab of the global view shows the information on sub-cohorts: (1.1a) list of involved patients (1.1b) Kaplan-Meier represents the overall survival of identified sub-cohorts and the whole cohort in the first tab. The second tab of the Kaplan-Meier plot shows the event-free survival information for the same (sub-)cohorts. (1.1c) represents the similarities based on the overlapped patients between sub-cohorts.

cos plots of the selected patients, see Fig. 5.6(1.3a). The expert can check the genomic abbreviation of a selected patient at all time points and from different sample tissues.

b. **Survival curves:** Another view shows the survival rates of identified sub-cohorts vs whole cohort using Kaplan-Meier plots. Kaplan-Meier is the most common and useful techniques to represent the ratio of patients who are still alive or did not experience a relapse (event-free survival) after a certain amount of time after diagnosis [90]. The Kaplan-Meier survival curve was proposed by Edward L. Kaplan and Paul Meier in 1958.

As shown in Figure 5.5(1.1b) , DiscoVA provides the overall survival(OS)/event-free survival (EFS) information of all sub-cohorts in two tabs. The x-axis shows the number of months after diagnosis, whereas the y-axis shows the survival/event-free probabilities. The survival function $S(t)$ for each sub-cohort is calculated by:

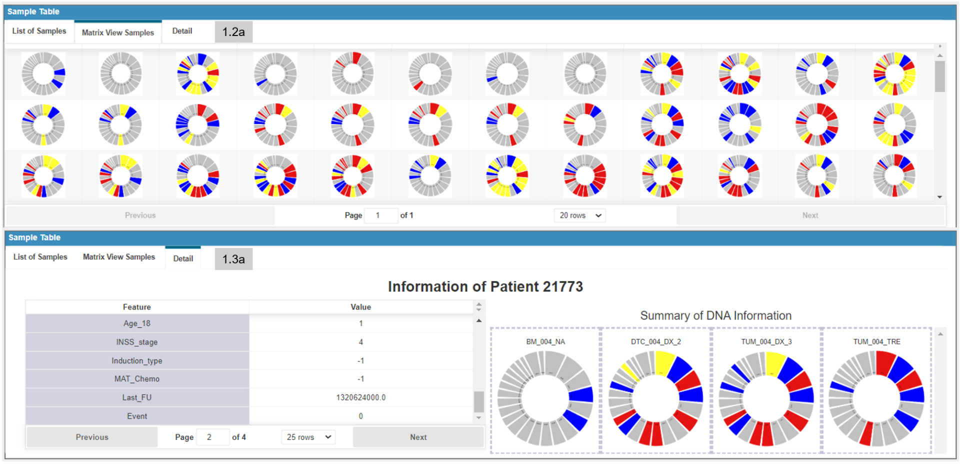
$$S(t) = \prod_{t_{(i)} \leq t} \frac{n_i - d_i}{n_i} = \prod_{t_{(i)} \leq t} \left(1 - \frac{d_i}{n_i}\right) \tag{5.2}$$

Where

$S(0) = 1$

$d_i$ = sample patients in which the event at the time $t_{(i)}$ has occurred

$n_i$ = sample patients at the time $t_{(i)}$ at risk



Information of Patient 21773

| Feature | Value |
|---|---|
| Age_18 | 1 |
| INSS_stage | 4 |
| Induction_type | -1 |
| MAT_Chemo | -1 |
| Last_FU | 1320624000.0 |
| Event | 0 |

Figure 5.6: (1.2a) The second tab of the overall view represents the list of samples of involved patients of the current sub-cohort. Thus, the expert can see an overview of losses and gains of all chromosome sets of involved patients in the selected sub-cohort. (1.3a) The third tab of the overall view shows the information of a selected patient.

c. **Similarity of sub-cohorts:** As shown in Figure 5.5(1.1c), the pairwise similarity of discovered sub-cohorts based on overlapped patients is shown using simple connected graphs, where each circle represents a sub-cohort and the radius size of the circle corresponds to the size of the sub-cohort (bigger radius, more cluster members). The size of circles is normalized appropriately. Moreover, to represent the overlaps, the thickness of the line between nodes (sub-cohorts) represents the amount of pairwise similarity (thicker line, more shared patients between the two sub-cohorts). The radius of circles is normalized appropriately to represent the number of involved patients within each sub-cohort. The closer sub-cohorts share more similar patients. The cosine similarity technique is used to calculate the pairwise similarity of sub-cohorts. This technique is popular in text mining to calculate the similarity of documents. In a work of [91], it is used to calculate the similarity between clinical trial cohorts. As represented in Equation 5.3, it measures the similarity between two non-zero vectors $A$ and $B$. With having more similarities between $A$ and $B$ vectors, the cosine similarity value is closer to 1. The cosine similarity matrix is fed to the multidimensional scaling (MDS) [92] technique to map each sub-cohort in 2D space.

Although there are possibilities to show the overlaps in plots, such as the Venn diagram, by having more sub-cohorts it will be cluttered or not intuitive to understand. Moreover, by Venn diagram encoding more information, like the size of the sub-cohort, is not possible. Alternatively, we model each cohort as a binary vector, with the length of the cohort's size, where the involved individuals are equal to 1 and non-involved ones are equal to zero. Then, we calculate the similarity between cohorts by cosine similarity, as it is appropriate in this situation and works well with high-dimensional data.

$$A \cdot B = \|A\| \, \|B\| \, cos\Theta \tag{5.3}$$

#### Clinical and MRD data section

This section provides inspection of clinical and MRD data. Additionally, the expert is enabled to specify a sub-cohort using linking and brushing. This section has three subsections (T1aa):

a. **MRD data view:** MRD data are shown in brushable parallel coordinates (PC), which show for each patient the number of cancer cells in a bone marrow sample in different phases of treatment. As shown in Figure 5.7(a), each y-axis represents a time point of treatment. The hovering of lines gives information on corresponding patients. The expert is allowed to select a group of patients by brushing at specific time points in any order.

b. **Clinical data view:** The same as with MRD data, the clinical information is shown in a PC plot. As the number of clinical features is high, i.e. 62 features, the user is enabled to add/remove the intended features for analysis from a drop-down box located on the left side of the panel. (Fig. 5.7(c)). Similar to the MRD data panel, the selection of a group

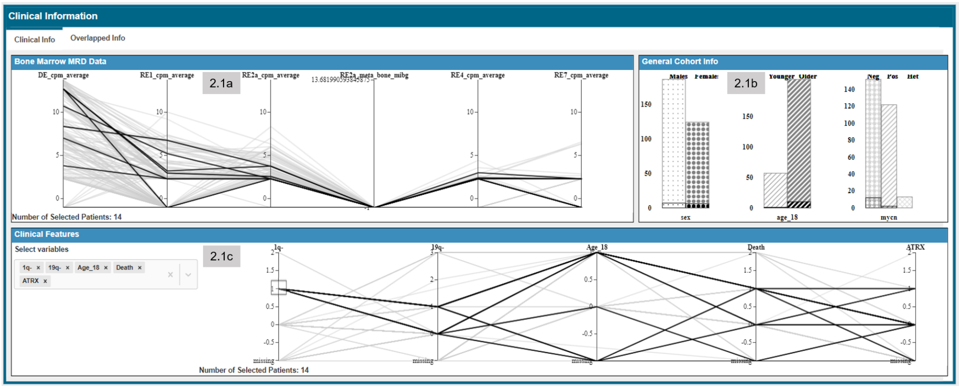


Figure 5.7: The first tab of clinical & MRD information consists of (a) the parallel coordinates of MRD data from different data points from the time of diagnosis until after treatment. (b) Bar charts to show proportions of the whole cohort vs. the selected sub-cohort of some specified features. (c) The parallel coordinates of the clinical data.

of patients is possible by brushing the patients at the desired intervals. Moreover, the tooltips give information about patients by hovering on the lines.

c. **Clinical data statistics:** As shown in Figure 5.7(b), the bar charts show the gender and known risk factors, i.e. features that are associated with a high risk of relapse or death. These features consist of sex, age at diagnosis, and *MYCN* amplifications). For example, based on previous studies by [93], patients younger than 18 months at the age of diagnosis have a better chance to be cured. Because all the colors are reserved to present sub-cohort features, in this view the bar charts are distinguished by different patterns.
As the user selects a sub-cohort, the proportion-intended bar chart will be highlighted to show the quantification of selected sub-cohorts.

### Genome info section

The genome section consists of CNA/DNA information of patients and is composed of three sub-sections (T1bb):

a. **Integrative Genome Viewer:** As presented in Figure 5.9(a), IGV as an interactive common visualization system for browsing genomic abbre-

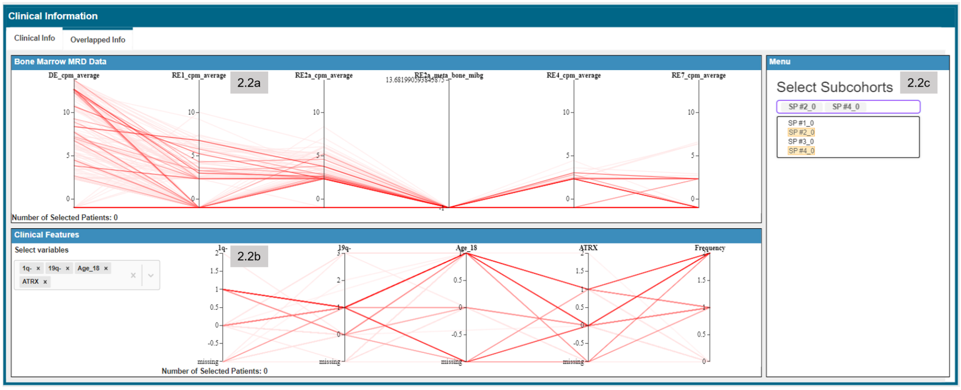
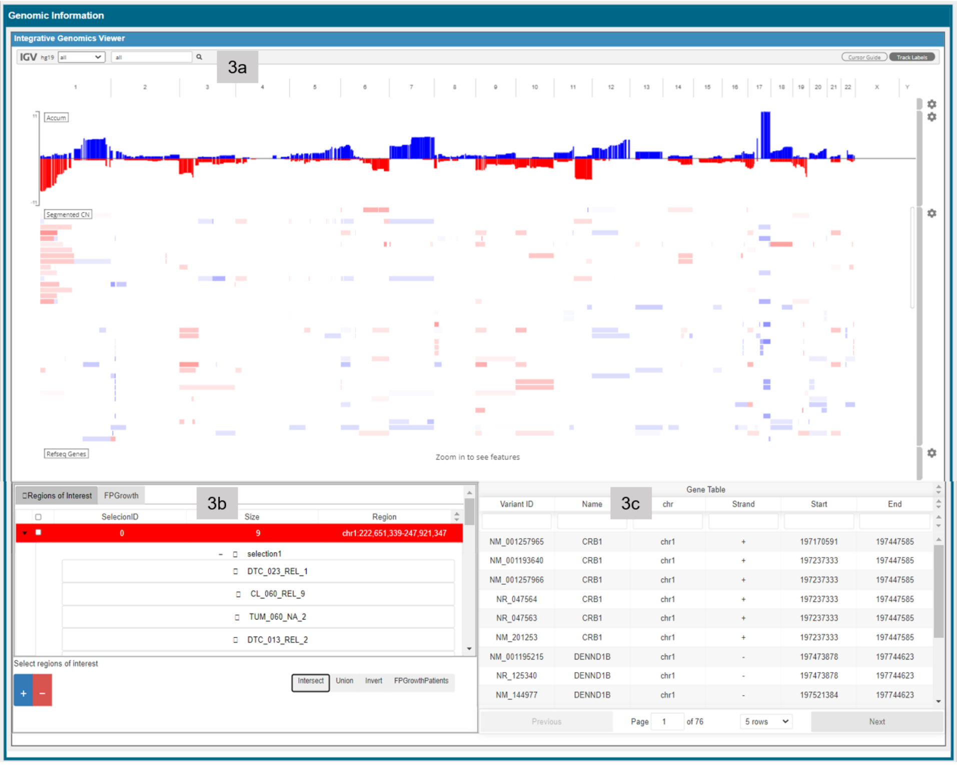


Figure 5.8: The second tab of clinical & MRD data is the overlapped selection, which enables the user to make queries based on the overlap between selected identified sub-cohorts.

viations is the main part of the genome section [94]. The embedded IGV is enhanced for this work and consists of four kinds of information, i.e. accumulative, aberrations (segment track), raw data, and gene tracks. In this enhanced IGV, the cumulative track is implemented to show the sum of genomics abbreviations in the chromosome interval. It helps the expert to understand the number of abnormalities in a selected sub-cohort and to identify the regions of interest. The expert can see the total frequency of deletions or gains as abnormalities in all chromosomes in a view, or zoom in to a specific chromosome.

The next segment track shows genomic abbreviations for each sample individually. Actually, each row shows the information of a specific sample, where the same color code as in circos plots is used as a convention, blue as gain, and red as losses. The color intensity of the segments represents the copy numbers in those chromosome intervals. As the expert selects the desired sample (a row), the raw data of the intended sample will be loaded to the raw data track.

b. **Genome query management:** This element is linked to the IGV and allows the user to make queries and select sub-cohorts based on the genome CNA information. This component is shown in Figure 5.9(b). To select a sub-cohort based on genome information, the expert uses the IGV browser and marks the region of interest using the query management component. For this purpose, based on the interestingness of



| ☐ | SelecionID | Size | Region |
|---|---|---|---|
| ☐ | 0 | 9 | chr1:222,651,339-247,921,347 |

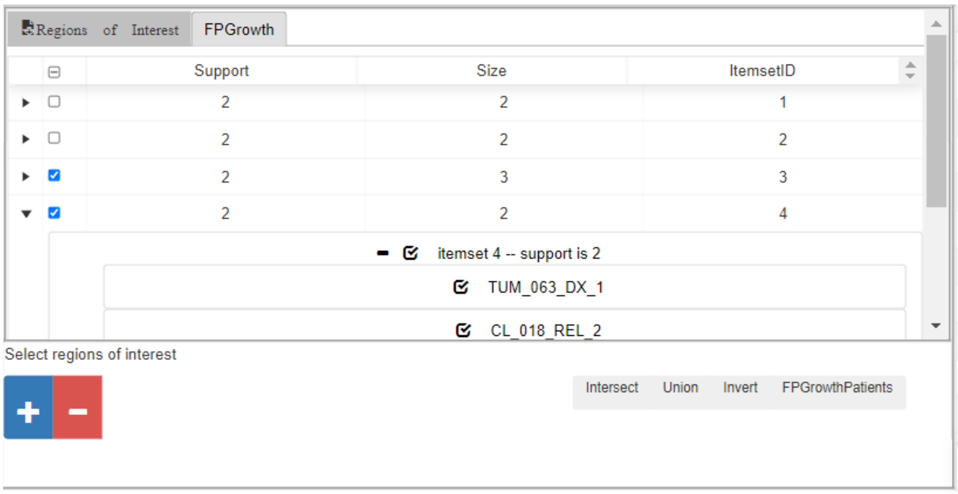
- ☐ selection1
  - ☐ DTC_023_REL_1
  - ☐ CL_060_REL_9
  - ☐ TUM_060_NA_2
  - ☐ DTC_013_REL_2

| Variant ID | Name | chr | Strand | Start | End |
|---|---|---|---|---|---|
| NM_001257965 | CRB1 | chr1 | + | 197170591 | 197447585 |
| NM_001193640 | CRB1 | chr1 | + | 197237333 | 197447585 |
| NM_001257966 | CRB1 | chr1 | + | 197237333 | 197447585 |
| NR_047564 | CRB1 | chr1 | + | 197237333 | 197447585 |
| NR_047563 | CRB1 | chr1 | + | 197237333 | 197447585 |
| NM_201253 | CRB1 | chr1 | + | 197237333 | 197447585 |
| NM_001195215 | DENND1B | chr1 | - | 197473878 | 197744623 |
| NR_125340 | DENND1B | chr1 | - | 197473878 | 197744623 |
| NM_144977 | DENND1B | chr1 | - | 197521384 | 197744623 |

Figure 5.9: The Genomic Information Panel comprises three components. The IGV (3a) represents the CNA information of all patients' samples and the raw data of the selected sample. Component 3b is attached to the IGV and is used for the management of regions of interest and settings to specify a sub-cohort. The table of genes (3c) represents the genes within selected locus intervals in IGV.

samples carrying out deletion or gains of copy numbers, the user saves their list by selecting the intended button (blue button to save gain and red button to save losses). Then, there are four options to pursue the query. First, getting the union of samples in regions of interest or make their intersection. Moreover, it is possible to invert the selection, i.e. select samples which are not located in the regions of interest.

Additionally, DiscoVA suggests the set of samples which frequently appeared together in the marked regions of interest using the FP-Growth algorithm [95]. As presented in Figure 5.10 FP-Growth itemsets are shown in the second tab of the query management panel and the expert can sort the sets based on the support value, which represents the frequency of co-appearance of samples, or order samples based on the



| | Support | Size | ItemsetID |
|---|---|---|---|
| ☐ | 2 | 2 | 1 |
| ☐ | 2 | 2 | 2 |
| ☑ | 2 | 3 | 3 |
| ☑ | 2 | 2 | 4 |

☑ itemset 4 -- support is 2
- ☑ TUM_063_DX_1
- ☑ CL_018_REL_2

Figure 5.10: DiscoVA uses the FP-Growth algorithm to find the set of samples which frequently appeared together in the regions of interest.

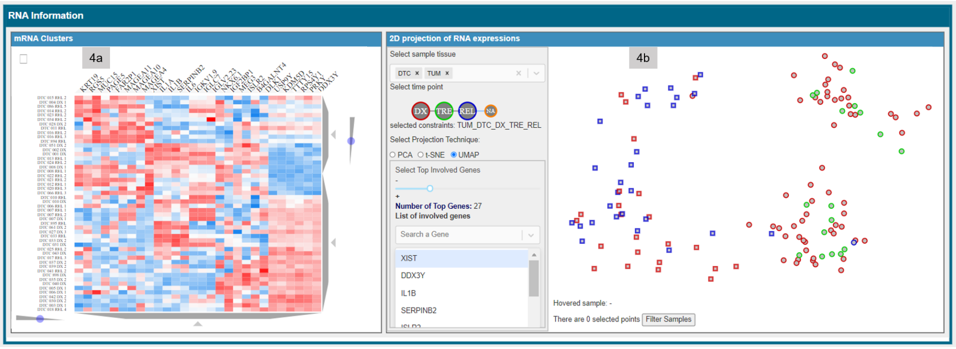
size of sets. Then, the expert can select the desired sets and filter the sub-cohort based on the samples in the selected regions.

c. **Genes table:** Giving information that is located in a region of interest is helpful for data analysis. The gene table is also linked to the IGV component and displays the list of genes and their information (variant ID, gene name, chromosome number, strand, and the start and end position of the genes ) of the currently selected chromosome interval of IGV. To get more information on the desired gene by selecting it, the user will be redirected to a page giving details about the gene from www.genecards.org repository.

#### mRNA expression data section

This section gives an overview of $x$ top high-variance mRNA expression data and consists of two sections (T1cc):

a. **Heatmap of mRNA clusters:** This panel visualizes the agglomerated hierarchical clusters [96] of mRNA data for the top selected genes in an interactive heatmap called clustergrammer [97], see Figure 5.11. The clusters are pre-calculated using the $stat$ package in R statistical tool [98].



Select sample tissue: DTC × TUM ×

Select time point

selected constraints: TUM_DTC_DX_TRE_REL

Select Projection Technique: ○ PCA ○ t-SNE ● UMAP

Select Top Involved Genes

**Number of Top Genes:** 27

**List of involved genes**

Search a Gene

Hovered sample: -

There are 0 selected points

Figure 5.11: The mRNA information consists of two main panels. (4a) The heatmap represents the clustered mRNA expression data for the top regulated genes. (4b) The second component shows the similarity between samples for the genes with highest variance. The distance between samples is calcuated by three techniques t-SNE, UMAP and PCA. Sub-cohorts can be filtered by brushing sample data points in this view.

b. **2D projection of mRNA expression data:** This panel shows the similarity of mRNA samples by projecting them in 2D space using a dimensionality reduction technique. As shown in Figure 5.11(b), the samples shown here can be filtered based on the tissue type (main tumor tissue or bone marrow sample) and different time points for the top $x$ high-variance gene (using the slider). The similarity of samples is precalculated employing the principal component analysis (PCA), t-SNE, and UMAP using python libraries.

The parameters of t-SNE including the number of iterations and the perplexity were set considering the total number of patients. We run t-SNE with 1000 iterations and perplexity was set to 20 as a feature for the number of the effective nearest neighborhoods regarding the size of our data.

The number of neighbors in UMAP is set to 15 and the minimum distance between points to 0.2. Moreover, the list of involved genes is shown in a list box. The top genes are selected based on their variance using the slider.

By a lasso brush, the user is allowed to select/filter a sub-cohort having any arbitrary shape based on mRNA expression data.

### Development

In this section, we describe the technologies and processes used for the development of the DiscoVA.

- **Implementation:** The development of DiscoVA is based on a client/server web interface. The backend server-side was implemented in the $node.js$ platform using the JavaScript programming language [99]. The front-end was implemented by React.js and Redux technologies.
  React.js, a JavaScript library developed by Facebook [100] allows building reusable components. Redux was used for state management. As JavaScript by itself is stateless, this technology enables the interaction between all components of the system.

- **Data format and pre-processing:** To be able to integrate and compare MRD data, collected at different time points and presenting highly heterogeneous data ranges, we did a logarithmic normalization to avoid data skew and to facilitate the interpretability of the plots.

  The missing values were filled by $NA$ or a negative distinct value. In addition, an R script [98] was used to transform the manually curated CNA information (.bed files) of each sample, to a single .seg file, in order to display genomic intervals and corresponding CNA information as separate tracks using the Integrative Genomics Viewer (IGV). All data files (RNA-, MRD- and clinical data) were provided as .csv files and were converted to *.json* files to be compatible with the system implementation requirements.

## 5.5 Evaluation

We evaluate the effectiveness of the system in two stages: first, the biomedical experts made a case study. The case study was base on a publication [85]. The aim of the case study was to allow the experts to validate an already confirmed hypothesis and let us know how easily they could reach it.

The team of biomedical experts was including three biologists (Fikret Rifatbegovic, Eva Bozasky, Sabine Taschner-Mandl) and a data scientist (Florian Kromp). Two participants of the expert team were involved in the design of the system (Sabine Taschner-Mandl and Florian Kromp). Participants were trained in a session on how to use the last version of the system. We made the case study in two separate sessions via screen sharing.

In the second stage, we provided a survey for the experts to rate DiscoVA w.r.t. three aspects: accuracy, visual analytics, and the interactions and visualization components.

### 5.5.1 Case Study

The case study is based on a publication [85]. In this case study, the team of experts developed strategies to investigate the prognostic relevance, and biological and clinical characteristics associated, with a particular genetic event that has potential relevance in the given cohort of neuroblastoma patients. [85] demonstrate that in bone marrow metastases certain markers involved in telomere maintenance, e.g. intragenic deletions of the ATRX gene, are frequently associated with copy number aberrations in the 1q and the 19q arm at the time point of relapse [85]. In other words, they found out that 1q- is highly associated with 1q deletion and ATRX deletion. Moreover, the 19q-[1] has an impact on event-free survival, but only on patients older than 18 months and non-MYCN amplified cases.

In the first part of the evaluation, we show how the user will be able to explore a single sub-cohort, and then we show how it is used to compare different identified sub-cohorts.

Thus, to make the hypothesis, the experts start with the IGV of CNAs information. From the embedded accumulative view they observed high frequencies in 1q-, 1q+, and 17q+. The experts were interested to get into detail with 1q- patients. Therefore, to zoom-in, they selected chromosome 1 from the IGV. They noticed that mainly samples with 1q- are located in the terminal $q$ region (end of the second half of chromosome one). Thus, they selected the terminal interval of $q$ region and from the genome query

---

[1] The first half of a chromosome is called $p$, and the second half is called $q$. Thus, for example, 19q- patients refer to patients who have a sample with at least a deletion (red segments in IGV) in the second half of their 19 chromosomes.

panel extracted and filtered the samples with deletion aberrations, see Fig. 5.9(3b). As a result of this step, all the corresponding samples/ patients were highlighted in all components.

In the next step, they scrolled up to clinical panel 5.7. It was noticed that the filtered 1q- CNA samples belong to five patients. The statistical bar charts showed that all the patients are male and older than 18 months, see Fig. 5.7(2.1b). Moreover, most of them are patients without *MYCN* amplification.

Then, they wanted to check the impact of the selected cohort on event-free survival curve 5.5(1.1b). Therefore, they referred to the global information panel. The event-free survival Kaplan Meier showed that the selected 1q- cohort has a poor event-free survival in comparison with the whole cohort. Also, the accumulative view of IGV shows with patients with 1q- it is more prominent than they have also deletions in the 19q chromosome.

As currently, genomics CNA data was available for only a few patients, then the experts decided to make the same query, but via genomics annotated data from clinical panel 5.7. Therefore, the saved the sub-cohort and reset the query for making a new query to compare patients with 1q- and without 1q-, i.e. have a deletion in the select half of chromosome 1 versus not having a segment with deletions in the second half of chromosome1, respectively.

The experts made a query by selecting the patients with 1q- using the parallel coordinates of the clinical panel following by selection of patients without 1q-, Fig. 5.7(2.1b), and saved the results of sub-cohorts. In addition, they wanted to compare the identified sub-cohorts with patients with a deletion in the *q* region of chromosome 19.

From the results (Fig. 5.5(1.1c)), they quickly determined that the sub-cohort of patients without 1q- has more members than others. It also shows obviously that it has no shared member with 1q- and *terminal 1q-* sub-cohorts. Next, by looking at the event-free survival Kaplan-Meier plot, the experts noticed that that the sub-cohorts of 1q- and 19q-, while sub-cohort of patients without 1q- have a better outcome, see Fig. 5.5(1.1b).

To more precisely evaluate DiscoVA, we asked the experts to fill a questionnaire about their experience using the system. In the following section, we discuss the results of the survey.

### 5.5.2 Survey

The purpose of the survey is to rate the system from the users' point according to measurements introduced in [101, 102] consisting of accuracy, analytic process, visualization, and interactions. We provided the users with 14 questions that cover the above measurements. The questions were designed in a rating style on a scale of 1 to 10 representative for completely dissatisfied and completely satisfied, respectively.

The first part of the questions measures the satisfaction of the users from an accuracy point. We listed questions on:

a. How intuitive are corresponding visualizations for each type of data to how the quantities of the data?

b. How good the layout is arranged, i.e. panels and sections?

c. How do you find the color choices for the visualizations and plots?

d. How appropriate are the visualizations for the type of data?

e. How do you rate the labeling and overall readability of the plots and data?

f. How do you rate the way that missing data was handles?

All the users rated the accuracy of the system between 8 and 10 which in descriptive words mean they were almost (completely) satisfied.

The second part of the questions was assessing the analytical process:

a. How do you rate the intuitiveness and clearness of the process flow (where to start and how to proceed)?

b. How good the system supports you with the analytical reasoning questions?

c. How the system support you to make queries faster than before?

All users agreed that DiscoVA fully supports the above-mentioned analytical process.

The third set of questions assesses the employed visualization and interactions:

a. How do you find functionalities of the system to display the desired data to help you explore the data, i.e. is it enough for the desired tasks?

b. How good and intuitive are views linked together to help you to make your desired query?

c. How consistent is the system?

d. How easy you learn and remember to use the system?

e. How familiar are you with the employed visualizations?

The result shows that the users feel comfortable using and interact with the system.

In general, the expert team was satisfied with DiscoVA. They believe that it makes the exploration of the data much faster than usual and independent of bioinformaticians. By using DiscoVA they can inspect the data in an integrative manner which was impossible before. However, they had suggestions to improve the clusters of mRNA data. It is ideal if the user could set of top genes for the clustering of mRNA data by selecting a chromosome interval from the IGV plot. However, to achieve this an online clustering should be provided which might be computationally intensive.

## 5.6 Summary and Conclusion

In this chapter, we described the developed system, DiscoVA, as a coordinated multiple view system for visual queries and the exploration of multi-omics datasets. The main aim of DiscoVA is to identify potential new prognostic biomarkers (features) and to build new hypotheses. DiscoVA allows the joint exploration of patient-related, clinical, transcriptomic, and genomic data of patient cohorts with cancer. We used the data of neuroblastoma patients as a rare pediatric cancer. Distinct sub-cohorts can be identified by brushing multiple linked datasets visualized in separate components. Parallel coordinates were employed to visualize clinical and other patient-related data. Sub-cohorts can be defined based on all views available. By using simple circos plots, genomic information of patient samples is summarized comprehensively.

Moreover, the IGV explorer was integrated and enhanced to summarize genomic copy number aberrations of a patient cohort and to select regions of interest to create new sub-cohorts. mRNA expression data was projected on a 2-dimensional space by applying the three-dimensionality reduction techniques PCA, t-SNE, and UMAP. The heatmap shows the hierarchy clusters of mRNA data for the selected top genes. The Kaplan-Meier plot shows event-free survival time and overall survival prediction for the saved sub-cohorts vs. the whole cohort. Visual analytics functions were applied to enable hierarchical sub-cohort management.

To evaluate DiscoVA, the system was used by biologists of Saint Anna Children's Hospital in Vienna, Austria. The biologists' feedback was generally very good. DiscoVA was considered to satisfy all requirements defined initially. Additional adaptations will be carried out to improve the design and allow a simultaneous view of all components. By implementing an unsupervised approach to identify discrete sub-cohorts, hidden relations within the dataset could be revealed. This is currently hampered by the limited number of samples/analyses available but will be possible in the future by constantly increasing the dataset. The existing components should be further improved by adding confidence intervals to the Kaplan-Meier plots.

# 6

# Conclusions

In this thesis, we investigated two types of data: epidemiological data and multi-omics data. The main aim of epidemiological studies is to reveal determinants of diseases beyond a specific population. Although the main purpose of epidemiology analysis is to define the risk factors of a specific disease, the data are collected from participants without considering if they have the disease or not.

In the multi-omics or integrative omics analysis, the data are carried out from multiple sources like mRNA and genome. The multi-omics data is usually from patients with a specific disease, e.g. cancer. Combining omics data from different sources enables biologists to inspect complex biological features and reveal the correlation between different omics features. The multi-omics analysis has a strong role in precision or individualized medicine and biomarker discovery (markers for diagnosis of specific diseases, e.g. cancer).

In Chapter 2, the history, terms, and the data mining technique for analysis of epidemiological studies is provided. First, an overview is given to introduce the people who had an important role in the evolution of epidemiological studies. Then, the type of data in such studies is described following the review of the source of the data. Generally, the data can be collected at a single point (cross-sectional) or at multiple points to observe the changes in the characteristics of the same participants.

Subsequently, the important data mining techniques (mostly the methods that are used in this thesis) can be used for the analysis of the large-scale and heterogeneous data, i.e. epidemiological and multi-omics data consisting of classification, clustering, dimension reduction, association rules, and feature selection are explained.

## 6.1 Visual Analytics of Missing Values

In Chapter 3, a description of data quality issues is given. The main focus of this chapter is on the missing values in epidemiological longitudinal studies, especially when there is a dependency for the missing values. Then, VIVID as a system for exploration and handling of missing values is introduced [42]. VIVID suggests multiple imputation for filling the missing values.

To evaluate VIVID in the terms of accuracy and time, we created a dummy dataset with MAR dependency for missing values. The result showed that the VIVID system can help to reduce the complexity of imputation by giving suggestions to the user, which do not affect the accuracy of imputations.

There are still suggestions to turn VIVID into a comprehensive framework for analysis and handling of missing data. Although the embedded visualization components in VIVID help expert users to interpret the missing values, there is still a need to boost them in terms of scalability by employing more interactions or zoom functions (e.g. the circular graph). Currently, VIVID can check the plausibility of each predicted future separately, while it can be helpful to check and compare the distribution of multiple features at the same time after the prediction process. VIVID considers that the type of missing values is MAR or MCAR, while MNAR is also a common case in epidemiological data. By the availability of additional external data and by the use of multiple imputation MNAR can also be handled. Additionally, to make suggestions for building the predictor matrix, VIVID only calculates the linear correlation between features, while involving the nonlinear relationships would be also an advantage. Furthermore, VIVID can be generalized in other fields like economy and politics by making an appropriate grouping of features (as most of the visualization components are color-coded based on the taxonomy of features).

## 6.2 Cross-Sectional & Longitudinal Epidemiological Data

Chapter 4 focuses on sub-population discovery in epidemiological data, mainly based on [103]. To tackle the issues with global clustering, we used subspace clustering to find clusters among cross-sectional epidemiological data. Then, we present S-ADVIsED as a prototype system for the exploration of the result of subspace clustering. As a result of subspace clustering needs to be transformed in a rectangular shape as well as validated, S-ADVIsED suggests a visual cluster description and classification technique to validate the results using another independent population.

DRESS subspace clustering that we used is a density-based clustering that finds clusters in arbitrary shapes. Furthermore, S-ADVIsED is only based on the analysis of the cross-sectional data. Thus as future work, mock-ups were provided to find groups of people in longitudinal epidemiological data using biclustering. It allows the analyst to explore the results of biclustering by following each cluster transition over time.

## 6.3 Visual Queries on Multi-Omics Data

In Chapter 5, DiscoVA as a web-based system for exploration and finding discriminate sub-cohorts of patients is presented. DiscoVA was specially designed and developed to identify biomarkers in cancer patients, especially neuroblastoma as rare cancer in childhood. The preliminary version of DiscoVA is published in [104]. The main advantage of DiscoVA is to let the analyst find sub-cohorts using different kinds of clinical and multi-omics data in an integrative and explorative approach. We have evaluated DiscoVA by making a case study and survey questionnaire. Overall, the expert team was satisfied with the usability of it and agreed that it makes all their analytical process much faster.

Besides all benefits of DiscoVA, its functionality can be improved by accomplishing the tasks from an explorative approach to a supervised approach. In the future, by the availability of more data, supervised approaches like Patient Similarity Networks (PSN) can be employed for automatic integration and consequently classification of different sources of the data. Moreover, the visualizations and interactions can be improved to ease the usage of DiscoVA, e.g. flexible arrangement and expansion of panels and adding confidence intervals to survival curves.

# Bibliography

[1] Nish Chaturvedi. Ethnicity as an epidemiological determinant—crudely racist or crucially important?, 2001.

[2] Kenneth J Rothman. *Epidemiology: an introduction*. Oxford university press, 2012.

[3] Bill Gates. Responding to covid-19—a once-in-a-century pandemic? *New England Journal of Medicine*, 2020.

[4] Kari S McLeod. Our sense of snow: the myth of john snow in medical geography. *Social science & medicine*, 50(7-8):923–935, 2000.

[5] Jae W Song and Kevin C Chung. Observational studies: cohort and case-control studies. *Plastic and reconstructive surgery*, 126(6):2234, 2010.

[6] Bernhard Preim and Kai Lawonn. A survey of visual analytics for public health. In *Computer Graphics Forum*, volume 39, pages 543–580. Wiley Online Library, 2020.

[7] Neil Pearce. Classification of epidemiological study designs. *International journal of epidemiology*, 41(2):393–397, 2012.

[8] Moyses Szklo and F Javier Nieto. *Epidemiology: beyond the basics*. Jones & Bartlett Publishers, 2014.

[9] Sotiris B Kotsiantis, I Zaharakis, and P Pintelas. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160(1):3–24, 2007.

[10] J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

[11] Rupali Bhardwaj and Sonia Vatta. Implementation of id3 algorithm. *International Journal of Advanced Research in Computer Science and Software Engineering*, 3(6), 2013.

[12] Joseph B Kruskal and James M Landwehr. Icicle plots: Better displays for hierarchical clustering. *The American Statistician*, 37(2):162–168, 1983.

[13] Stef Van Den Elzen and Jarke J Van Wijk. Baobabview: Interactive construction and analysis of decision trees. In *2011 IEEE conference on visual analytics science and technology (VAST)*, pages 151–160. IEEE, 2011.

[14] Guoqiang Peter Zhang. Neural networks for classification: a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 30(4):451–462, 2000.

[15] Sotiris Kotsiantis and Dimitris Kanellopoulos. Association rules mining: A recent overview. *GESTS International Transactions on Computer Science and Engineering*, 32(1):71–82, 2006.

[16] Michael Hahsler and Sudheer Chelluboina. Visualizing association rules: Introduction to the r-extension package arulesviz. *R project module*, pages 223–238, 2011.

[17] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499, 1994.

[18] Jiawei Han, Jian Pei, and Yiwen Yin. Mining frequent patterns without candidate generation. *ACM sigmod record*, 29(2):1–12, 2000.

[19] Chen Meng, Oana A Zeleznik, Gerhard G Thallinger, Bernhard Kuster, Amin M Gholami, and Aedín C Culhane. Dimension reduction techniques for the integrative analysis of multi-omics data. *Briefings in bioinformatics*, 17(4):628–641, 2016.

[20] Bum Chul Kwon, Ben Eysenbach, Janu Verma, Kenney Ng, Christopher De Filippi, Walter F Stewart, and Adam Perer. Clustervision: Visual supervision of unsupervised clustering. *IEEE transactions on visualization and computer graphics*, 24(1):142–151, 2017.

[21] Lance Parsons, Ehtesham Haque, and Huan Liu. Subspace clustering for high dimensional data: a review. *ACM SIGKDD Explorations Newsletter*, 6(1):90–105, 2004.

[22] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

[23] Michael AA Cox and Trevor F Cox. Multidimensional scaling. In *Handbook of data visualization*, pages 315–347. Springer, 2008.

[24] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[25] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[26] Aristidis Likas, Nikos Vlassis, and Jakob J Verbeek. The global k-means clustering algorithm. *Pattern recognition*, 36(2):451–461, 2003.

[27] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.

[28] Kamran Khan, Saif Ur Rehman, Kamran Aziz, Simon Fong, and Sababady Sarasvady. Dbscan: Past, present and future. In *The fifth international conference on the applications of digital information and web technologies (ICADIWT 2014)*, pages 232–238. IEEE, 2014.

[29] Fionn Murtagh and Pedro Contreras. Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(1):86–97, 2012.

[30] Juan Xie, Anjun Ma, Anne Fennell, Qin Ma, and Jing Zhao. It is time to apply biclustering: a comprehensive review of biclustering applications in biological and biomedical data. *Briefings in bioinformatics*, 20(4):1450–1465, 2019.

[31] Sara C Madeira and Arlindo L Oliveira. Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM transactions on computational biology and bioinformatics*, 1(1):24–45, 2004.

[32] M Arfan Ikram, Guy GO Brusselle, Sarwa Darwish Murad, Cornelia M van Duijn, Oscar H Franco, André Goedegebure, Caroline CW Klaver, Tamar EC Nijsten, Robin P Peeters, Bruno H Stricker, et al. The rotterdam study: 2018 update on objectives, design and main results. *European journal of epidemiology*, 32(9):807–850, 2017.

[33] Henry Völzke, Dietrich Alte, Carsten Oliver Schmidt, Dörte Radke, Roberto Lorbeer, Nele Friedrich, Nicole Aumann, Katharina Lau, Michael Piontek, Gabriele Born, et al. Cohort profile: the study of health in pomerania. *International journal of epidemiology*, 40(2):294–307, 2011.

[34] Theresia Gschwandtner, Johannes Gärtner, Wolfgang Aigner, and Silvia Miksch. A taxonomy of dirty time-oriented data. In *International Conference on Availability, Reliability, and Security*, pages 58–72. Springer, 2012.

[35] Won Kim, Byoung-Ju Choi, Eui-Kyeong Hong, Soo-Kyung Kim, and Doheon Lee. A taxonomy of dirty data. *Data mining and knowledge discovery*, 7(1):81–99, 2003.

[36] Hyun Kang. The prevention and handling of the missing data. *Korean journal of anesthesiology*, 64(5):402, 2013.

[37] Stef Van Buuren. *Flexible imputation of missing data*. Chapman and Hall/CRC, 2018.

[38] James Honaker, Gary King, Matthew Blackwell, et al. Amelia ii: A program for missing data. *Journal of statistical software*, 45(7):1–47, 2011.

[39] Sara Johansson Fernstad and Robert C Glen. Visual analysis of missing data—to see what isn't there. In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 249–250. IEEE, 2014.

[40] Patrick Royston. Multiple imputation of missing values. *The Stata Journal*, 4(3):227–241, 2004.

[41] Ian R White, Patrick Royston, and Angela M Wood. Multiple imputation using chained equations: issues and guidance for practice. *Statistics in medicine*, 30(4):377–399, 2011.

[42] Shiva Alemzadeh, Uli Niemann, Till Ittermann, Henry Völzke, Daniel Schneider, Myra Spiliopoulou, Katja Bühler, and Bernhard Preim. Visual analysis of missing values in longitudinal cohort study data. In *Computer Graphics Forum*, volume 39, pages 63–75. Wiley Online Library, 2020.

[43] Matthias Templ, Andreas Alfons, and Peter Filzmoser. Exploring incomplete data using visualization techniques. *Advances in Data Analysis and Classification*, 6(1):29–47, 2012.

[44] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.

[45] Kurt Hornik, David Meyer Karatzoglou, Achim Zeileis, and Maintainer Kurt Hornik. The RWeka package. 2007.

[46] Todd E Bodner. What improves with increased missing data imputations? *Structural Equation Modeling*, 15(4):651–675, 2008.

[47] John W Graham, Allison E Olchowski, and Tamika D Gilreath. How many imputations are really needed? some practical clarifications of multiple imputation theory. *Prevention Science*, 8(3):206–213, 2007.

[48] Linda M Collins, Joseph L Schafer, and Chi-Ming Kam. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, 6(4):330–351, 2001.

[49] T Alan Keahey. Using visualization to understand big data. *IBM Business Analytics Advanced Visualisation*, 2013.

[50] S van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, pages 1–68, 2010.

[51] Peter Kampstra. Beanplot: A boxplot alternative for visual comparison of distributions. *Journal of Statistical Software*, 28(1):1–9, 2008.

[52] Karin Biering, Niels Henrik Hjollund, and Morten Frydenberg. Using multiple imputation to deal with missing data and attrition in longitudinal studies with repeated measures of patient-reported outcomes. *Clinical Epidemiology*, 7:91–106, 2015.

[53] JM Antó and P Cullinan. Clusters, classification and epidemiology of interstitial lung diseases: concepts, methods and critical reflections. *European Respiratory Journal*, 18(32 suppl):101s–106s, 2001.

[54] Tommy Hielscher, Myra Spiliopoulou, et al. Identifying relevant features for a multi-factorial disorder with constraint-based subspace clustering. In *Computer-Based Medical Systems (CBMS), IEEE 29th International Symposium on*, pages 207–212, 2016.

[55] Shiva Alemzadeh, Tommy Hielscher, Uli Niemann, Lena Cibulski, Till Ittermann, Henry Völzke, Myra Spiliopoulou, and Bernhard Preim. Subpopulation Discovery and Validation in Epidemiological Data. In *EuroVis Workshop on Visual Analytics*, 2017.

[56] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. pages 226–231. AAAI Press, 1996.

[57] U. Niemann, T. Hielscher, M. Spiliopoulou, H. Völzke, and J. P. Kühn. Can we classify the participants of a longitudinal epidemiological study from their previous evolution? In *Proc. of IEEE Symposium on Computer-Based Medical Systems*, pages 121–126, 2015.

[58] Elke Achtert, Christian Böhm, Hans-Peter Kriegel, Peer Kröger, Ina Müller-Gorman, and Arthur Zimek. Detection and visualization of subspace cluster hierarchies. In *Proc. of Advances in Databases: Concepts, Systems and Applications, DASFAA*, pages 152–163, 2007.

[59] Ira Assent, Ralph Krieger, Emmanuel Müller, and Thomas Seidl. Visa: Visual subspace clustering analysis. *SIGKDD Explor. Newsl.*, 9(2):5–12, December 2007.

[60] Andrada Tatu, Leishi Zhang, Enrico Bertini, Tobias Schreck, Daniel Keim, Sebastian Bremm, and Tatiana Von Landesberger. Clustnails: Visual analysis of subspace clusters. *Tsinghua Science and Technology*, 17(4):419–428, 2012.

[61] Michael Hund, Dominic Böhm, Werner Sturm, Michael Sedlmair, Tobias Schreck, Torsten Ullrich, Daniel A Keim, Ljiljana Majnaric, and Andreas Holzinger. Visual analytics for concept exploration in subspaces of patient groups. *Brain Informatics*, pages 1–15, 2016.

[62] Jan Burmeister, Jürgen Bernard, Thorsten May, and Jörn Kohlhammer. Self-service data preprocessing and cohort analysis for medical

researchers. In *2019 IEEE Workshop on Visual Analytics in Healthcare (VAHC)*, pages 17–24. IEEE, 2019.

[63] Danlu Liu, William Baskett, David Beversdorf, and Chi-Ren Shyu. Exploratory data mining for subgroup cohort discoveries and prioritization. *IEEE journal of biomedical and health informatics*, 2019.

[64] Henry Völzke. Study of health in pomerania (ship). *Bundesgesundheitsblatt-Gesundheitsforschung-Gesundheitsschutz*, 55(6-7):790–794, 2012.

[65] Bernhard Preim, Paul Klemm, Helwig Hauser, Katrin Hegenscheid, Steffen Oeltze, Klaus Toennies, and Henry Völzke. Visual analytics of image-centric cohort studies in epidemiology. In *Visualization in Medicine and Life Sciences III*, pages 221–248. Springer, 2016.

[66] Daniel E Ho, Kosuke Imai, Gary King, and Elizabeth A Stuart. Matchit: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, 42(8):1–28, 2011.

[67] Joseph B Kruskal and Myron Wish. *Multidimensional scaling*, volume 11. Sage, 1978.

[68] Lena Cibulski. Visual analytics support for analysis of cohort study data: Requirements and concepts. *Project report, Otto-Von-Guericke University Magdeburg*, 2016.

[69] Edward WJ Curry. A framework for generalized subspace pattern mining in high-dimensional datasets. *BMC bioinformatics*, 15(1):1–14, 2014.

[70] Barbara C Worst, Cornelis M van Tilburg, Gnana Prakash Balasubramanian, Petra Fiesel, Ruth Witt, Angelika Freitag, Miream Boudalil, Christopher Previti, Stephan Wolf, Sabine Schmidt, et al. Next-generation personalised medicine for high-risk paediatric cancer patients–the inform pilot study. *European Journal of Cancer*, 65:91–101, 2016.

[71] Jinfeng Zou and Edwin Wang. Cancer biomarker discovery for precision medicine: New progress. *Current medicinal chemistry*, 2019.

[72] Felix Sahm, Daniel Schrimpf, Damian Stichel, David TW Jones, Thomas Hielscher, Sebastian Schefzyk, Konstantin Okonechnikov, Christian Koelsche, David E Reuss, David Capper, et al. Dna methylation-based classification and grading system for meningioma: a multicentre, retrospective analysis. *The Lancet Oncology*, 18(5):682–694, 2017.

[73] Jacques JM van Dongen, Taku Seriu, E Renate Panzer-Grümayer, Andrea Biondi, Marja J Pongers-Willemse, Lilly Corral, Frank Stolz, Martin Schrappe, Giuseppe Masera, Willem A Kamps, et al. Prognostic value of minimal residual disease in acute lymphoblastic leukaemia in childhood. *The Lancet*, 352(9142):1731–1738, 1998.

[74] Gemma Gatta, Riccardo Capocaccia, Laura Botta, Sandra Mallone, Roberta De Angelis, Eva Ardanaz, Harry Comber, Nadya Dimitrova, Maarit K Leinonen, Sabine Siesling, et al. Burden and centralised treatment in europe of rare tumours: results of rarecarenet—a population-based study. *The Lancet Oncology*, 18(8):1022–1039, 2017.

[75] Raveendran K Pillai and K Jayasree. Rare cancers: challenges & issues. *The Indian journal of medical research*, 145(1):17, 2017.

[76] Gemma Gatta, Jan Maarten Van Der Zwan, Paolo G Casali, Sabine Siesling, Angelo Paolo Dei Tos, Ian Kunkler, Renée Otter, Lisa Licitra, Sandra Mallone, Andrea Tavilla, et al. Rare cancers are not so rare: the rare cancer burden in europe. *European journal of cancer*, 47(17):2493–2511, 2011.

[77] Nicola Keat, Kate Law, Andrea McConnell, Matthew Seymour, Jack Welch, Ted Trimble, Denis Lacombe, and Anastassia Negrouk. International rare cancers initiative (irci). *ecancermedicalscience*, 7, 2013.

[78] Gudrun Schleiermacher, Isabelle Janoueix-Lerosey, and Olivier Delattre. Recent insights into the biology of neuroblastoma. *International journal of cancer*, 135(10):2249–2261, 2014.

[79] Ruth Ladenstein, Ulrike Pötschger, Andrew DJ Pearson, Penelope Brock, Roberto Luksch, Victoria Castel, Isaac Yaniv, Vassilios Papadakis, Geneviève Laureys, Josef Malis, et al. Busulfan and melphalan versus carboplatin, etoposide, and melphalan as high-dose

chemotherapy for high-risk neuroblastoma (hr-nbl1/siopen): an international, randomised, multi-arm, open-label, phase 3 trial. *The lancet oncology*, 18(4):500–514, 2017.

[80] Fikret Rifatbegovic, Christian Frech, M Reza Abbasi, Sabine Taschner-Mandl, Tamara Weiss, Wolfgang M Schmidt, Iris Schmidt, Ruth Ladenstein, Inge M Ambros, and Peter F Ambros. Neuroblastoma cells undergo transcriptomic alterations upon dissemination into the bone marrow and subsequent tumor progression. *International journal of cancer*, 142(2):297–307, 2018.

[81] G Elisabeta Marai, Chihua Ma, Andrew Thomas Burks, Filippo Pellolio, Guadalupe Canahuate, David M Vock, Abdallah SR Mohamed, and Clifton David Fuller. Precision risk analysis of cancer therapy with interactive nomograms and survival plots. *IEEE transactions on visualization and computer graphics*, 25(4):1732–1745, 2019.

[82] Daniel Bottomly, Shannon K McWeeney, and Beth Wilmot. Hitwalker2: visual analytics for precision medicine and beyond. *Bioinformatics*, 32(8):1253–1255, 2015.

[83] Alexander Lex, Marc Streit, H-J Schulz, Christian Partl, Dieter Schmalstieg, Peter J Park, and Nils Gehlenborg. Stratomex: visual analysis of large-scale heterogeneous genomics data for cancer subtype characterization. In *Computer graphics forum*, volume 31, pages 1175–1184. Wiley Online Library, 2012.

[84] Marc Streit, Alexander Lex, Samuel Gratzl, Christian Partl, Dieter Schmalstieg, Hanspeter Pfister, Peter J Park, and Nils Gehlenborg. Guided visual exploration of genomic stratifications in cancer. *Nature methods*, 11(9):884, 2014.

[85] M Reza Abbasi, Fikret Rifatbegovic, Clemens Brunner, Georg Mann, Andrea Ziegler, Ulrike Pötschger, Roman Crazzolara, Marek Ussowicz, Martin Benesch, Georg Ebetsberger-Dachs, et al. Impact of disseminated neuroblastoma cells on the identification of the relapse-seeding clone. *Clinical Cancer Research*, 23(15):4224–4232, 2017.

[86] André Oberthuer, Dilafruz Juraeva, Barbara Hero, Ruth Volland, Carolina Sterz, Rene Schmidt, Andreas Faldum, Yvonne Kahlert,

Anne Engesser, Shahab Asgharzadeh, et al. Revised risk estimation and treatment stratification of low-and intermediate-risk neuroblastoma patients by integrating clinical and molecular prognostic markers. *Clinical Cancer Research*, 21(8):1904–1915, 2015.

[87] Sandra Ackermann, Maria Cartolano, Barbara Hero, Anne Welte, Yvonne Kahlert, Andrea Roderwieser, Christoph Bartenhagen, Esther Walter, Judith Gecht, Laura Kerschke, et al. A mechanistic classification of clinical phenotypes in neuroblastoma. *Science*, 362(6419):1165–1170, 2018.

[88] Gabor Méhes, Andrea Luegmayr, Rosa Kornmüller, Ingeborg M Ambros, Ruth Ladenstein, Helmut Gadner, and Peter F Ambros. Detection of disseminated tumor cells in neuroblastoma: 3 log improvement in sensitivity by automatic immunofluorescence plus fish (aipf) analysis compared with classical bone marrow cytology. *The American journal of pathology*, 163(2):393–399, 2003.

[89] Tamasz Szczepariski, Alberto Orfão, Vincent HJ van der Valden, Jésus F San Miguel, and Jacques JM van Dongen. Minimal residual disease in leukaemia patients. *The lancet oncology*, 2(7):409–417, 2001.

[90] Manish Kumar Goel, Pardeep Khanna, and Jugal Kishore. Understanding survival analysis: Kaplan-meier estimate. *International journal of Ayurveda research*, 1(4):274, 2010.

[91] Matthew C Lenert, Dara E Mize, and Colin G Walsh. X marks the spot: Mapping similarity between clinical trial cohorts and us counties. In *AMIA Annual Symposium Proceedings*, volume 2017, page 1110. American Medical Informatics Association, 2017.

[92] Trevor F Cox and Michael AA Cox. *Multidimensional scaling*. Chapman and hall/CRC, 2000.

[93] Wen-Guang He, Yu Yan, Wen Tang, Rong Cai, and Gang Ren. Clinical and biological features of neuroblastic tumors: A comparison of neuroblastoma and ganglioneuroblastoma. *Oncotarget*, 8(23):37730, 2017.

[94] Helga Thorvaldsdóttir, James T Robinson, and Jill P Mesirov. Integrative genomics viewer (igv): high-performance genomics data visualization and exploration. *Briefings in bioinformatics*, 14(2):178–192, 2013.

[95] S Pramod and OP Vyas. Survey on frequent item set mining algorithms. *International journal of computer applications*, 1(15):86–91, 2010.

[96] Fionn Murtagh and Pierre Legendre. Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion? *Journal of classification*, 31(3):274–295, 2014.

[97] Nicolas F Fernandez, Gregory W Gundersen, Adeeb Rahman, Mark L Grimes, Klarisa Rikova, Peter Hornbeck, and Avi Ma'ayan. Clustergrammer, a web-based heatmap visualization and analysis tool for high-dimensional biological data. *Scientific data*, 4:170151, 2017.

[98] R Core Team et al. R: A language and environment for statistical computing. 2013.

[99] Stefan Tilkov and Steve Vinoski. Node. js: Using javascript to build high-performance network programs. *IEEE Internet Computing*, 14(6):80–83, 2010.

[100] Cory Gackenheimer. What is react? In *Introduction to React*, pages 1–20. Springer, 2015.

[101] Jean Scholtz, Catherine Plaisant, Mark Whiting, and Georges Grinstein. Evaluation of visual analytics environments: The road to the visual analytics science and technology challenge evaluation methodology. *Information Visualization*, 13(4):326–335, 2014.

[102] Kristin A Cook, Jean Scholtz, and Mark A Whiting. A software developer's guide to informal evaluation of visual analytics environments using vast challenge information. In *2015 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 193–194. IEEE, 2015.

[103] Shiva Alemzadeh, Uli Niemann, Till Ittermann, Henry Völzke, Daniel Schneider, Myra Spiliopoulou, and Bernhard Preim. Visual analytics of missing data in epidemiological cohort studies. In *VCBM*, pages 43–51, 2017.

[104] Shiva Alemzadeh, Florian Kromp, Bernhard Preim, Sabine Taschner-Mandl, and Katja Bühler. A visual analytics approach for patient stratification and biomarker discovery. In *VCBM*, pages 91–95, 2019.

# Declaration of Academic Integrity

I hereby declare that I have written the present work myself and did not use any sources or tools other than the ones indicated.

Datum: ....................................................................

(Signature)